Université d'Angers
Habilitation à Diriger des Recherches

Problèmes inverses d'apprentissage, apprentissage en ligne et applications

Sébastien Loustau

Document présenté le 2 décembre 2014

Rapporteurs

Stéphane Boucheron
Gábor Lugosi
Sara van de Geer

Composition du jury

Jérémie Bigot
Stéphane Boucheron
Loïc Chaumont
Fabienne Comte
Arnak Dalalyan
Aurélien Garivier
Laurence Hibrand-Saint Oyant

– Ce manuscrit est dédié à la mémoire de Laurent Cavalier,
disparu soudainement au début de l'année 2014. –

Remerciements

Je tiens tout d'abord à remercier et rendre hommage à mon directeur de thèse, Laurent Cavalier, qui a disparu subitement au début de l'année et qui est à l'origine de tout ce qui suit. Merci de m'avoir lancé sur ce vaste sujet à la frontière entre statistique mathématique et apprentissage. Merci de m'avoir fait confiance et de m'avoir laissé "grandir".

Je suis très heureux que Stéphane Boucheron, Gábor Lugosi et Sara van de Geer aient accepté de rapporter ce mémoire d'habilitation. Je les remercie du temps passé à la relecture du manuscrit, et de la qualité de leurs rapports (avec une mention spéciale pour Stéphane et son rapport de 5 pages). Je tiens ensuite à remercier tous les membres du jury d'avoir accepté de participer à la soutenance.

J'aimerai aussi remercier tous les membres du Larema pour leur sympathie. L'acclimatation à la douceur angevine (sic) a été facilitée par l'atmosphère qui règne dans ce département. Un grand merci au secrétariat pour la ponctualité des pauses cafés, et plus largement, à tout le personnel administratif de l'université, qui m'a souvent aiguillé dans les méandres administratifs.

Le quotidien de ces 5 années de maître de conférence a été rythmé par de nombreuses collaborations, ou simples discussions avec de nombreux collègues que je voudrai remercier ici. Je voudrais remercier particulièrement Clément Marteau et Michaël Chichignoud, ancien collègue de thèse pour ces 5 années de trouvailles dans la bonne humeur, entre Toulouse, Zurich, Bruxelles et Angers. Merci plus largement à tous mes co-auteurs Camille Brunet, Simon Souchet, Koji Kawamura, Laurence Hibrand et toute l'équipe Genhort de l'I.N.R.A. Dans le petit monde des stats maths, je voulais aussi remercier Arnak Dalalyan, Benjamin Guedj, Sébastien Gerchinovitz, Guillaume Lecué, Sacha Tsybakov, Joseph Salmon, Pierre Alquier, Bertrand Michel, et tout ceux que j'oublie pour ces invitations/discussions/conseils autour d'un séminaire/théorème/verre. Mon "isolement" à Angers n'a été possible que grâce à vous !

Je tiens à exprimer toute ma reconnaissance à ma famille pour son accompagnement dans ce long parcours, de Pau à Angers en passant par Marseille. La distance qui nous sépare depuis 10 ans diminuera bien un jour, après cette Habilitation à D. R. (merci Pierre).

Je tiens enfin à remercier tout particulièrement ma petite famille qui s'est agrandie dès mon arrivée à Angers : merci à Lucie et Adrien pour votre joie de vivre et un grand merci à Sarah pour la confiance dont elle me témoigne en me suivant dans cette aventure, cette fois-ci dans le grand Nord.

Préambule

Ce manuscrit est un recueil des différents travaux de recherche effectués depuis ma nomination à Angers il y a 5 ans. L'objectif est de tenter d'obtenir, autant que possible, un ensemble cohérent et équilibré entre résultats théoriques et applications. Le point de départ de ce manuscrit est purement théorique :

Quelle est la vitesse minimax en classification avec erreurs dans les variables ?

Cette question, posée presque innocemment par Clément Marteau aux rencontres de statistiques mathématiques du C.I.R.M. en 2009, nous a occupé près de 2 années entières, pour une réponse partielle dans [L3] et [L8]. En deux mots : hypothèse de marge, vitesses rapides et problèmes inverses ne font pas forcément bon ménage. De fil en aiguille, la généralisation vers d'autres problèmes d'apprentissage ([L4]), puis le passage au cas non-supervisé ([L10], [L16]) ont permis l'écriture d'un algorithme pour résoudre le problème de segmentation (ou clustering) de données bruitées. Restait la partie programmation, que j'ai confié à Camille Brunet ([L10]), puis à Simon Souchet ([L12]), grâce à un premier contrat de valorisation. Cette belle histoire est avant tout une aventure humaine : de la recherche de la bonne famille d'hypothèses vérifiant le triumvirat marge-régularité-haute fréquence avec Clément Marteau, ancien collègue de thèse, aux discussions algorithmiques avec Camille Brunet, qui a mis au point une version Beta de l'algorithme en à peine 4 mois.

A ce stade, nous sommes à la fin du Chapitre 2, et nous n'avons pas encore résolu le problème majeur de notre algorithme : *le choix de la fenêtre*. Face à un problème avec erreurs dans les variables, un excès de risque, et un compromis biais-variance à l'origine des bornes supérieures, la méthode de Lepski nous a paru idoine. En collaboration avec Michaël Chichignoud, l'autre ancien collègue phocéen, nous proposons une méthode adaptative qui utilise l'heuristique de Lepski et la comparaison d'estimateurs ([1]). Dans notre cadre, les estimateurs sont des risques empiriques et les résultats sont des vitesses rapides adaptatives pour l'excès de risque. Et puis les choses s'accélèrent. Pour traiter le cas des fonctions anisotropes, il faut considérer des ensembles de fenêtres plus vastes. Dans le cadre du bruit blanc ou de l'estimation de densité, on connaît la méthode de Goldenshluger et Lepski, mais son application directe aux cas des vitesses rapides n'est pas immédiate. L'idée est alors de changer de critère, et de considérer le gradient du risque pour mesurer la performance des estimateurs. Tout l'intérêt de ce critère réside en une phrase : une vitesse lente pour le gradient abouti à une vitesse rapide pour l'excès de risque, à condition que la perte soit suffisamment lisse ([L7]). On obtient donc *in fine* des résultats adaptatifs optimaux dans le cadre anisotrope en clustering, et comme corollaire immédiat, les premières vitesses minimax adaptatives en norme L_p pour des estimateurs non-linéaires dans [L7].

Les résultats des Chapitre 2 et 3 forment une première contribution du minimax à l'algorithme du problème de classification à partir de données indirectes. Il reste de nombreuses interrogations, qui seront abordées dans le futur, pour qui s'intéresse à ce sujet très vaste. La fin de ce mémoire liste quelques problèmes ouverts, qui découlent de l'écriture de cette partie du mémoire.

Début septembre 2013, en obtenant un congés pour recherche, je me suis dirigé vers une autre thématique : la prévision de suites individuelles. Là encore, c'est une rencontre, celle de Sébastien Gerchinovitz à Nantes, qui m'a convaincu que ce sujet valait la peine d'être exploré. Le dernier chapitre théorique de cette habilitation propose donc un premier tour d'horizon de ma modeste contribution dans ce domaine. La question posée est la suivante :

Peut-on proposer un cadre non-supervisé à la prédiction de suites individuelles ?

Autrement dit, que se passe-t'il dans le problème de prédiction de suites individuelles quand on n'a pas accès à un ensemble d'experts ? Cette question m'a vite dirigé vers les résultats de Jean-Yves Audibert qui propose de traiter de manière unifiée les problèmes en ligne et les problèmes statistiques traditionnels (c'est-à-dire i.i.d. ou batch). La simple application de ces résultats, et l'introduction de lois a priori particulières issus de la statistique bayésienne en grande dimension m'ont permis d'obtenir des bornes de regrets pour un algorithme séquentiel de clustering en ligne ([L9]). Cet algorithme n'a besoin d'aucune connaissance a priori sur le nombres de classes, ni sur des avis d'éventuels experts. Cette nouvelle direction m'a occupé ces dernièrs mois : l'obtention d'un algorithme complètement automatique, ainsi qu'une borne inférieure sur le regret, sont aussi étudiés dans ce manuscrit. L'extension à un cadre de bi-clustering ([L11]) permet également un nouveau regard sur le problème de complétion de matrices et la construction de systèmes de recommandations en ligne.

Enfin, ma nomination à Angers m'a permis de nouer de multiples contacts avec des chercheurs d'autres disciplines. Ces collaborations sont résumées dans le dernier chapitre de ce manuscrit, où l'on s'éloigne un peu des statistiques mathématiques pour résoudre des problèmes biologiques du vivant. La recherche de QTL (Quantitative Trait Locus) dans une population de rosiers ([L5], [L17]) est le premier sujet que j'ai abordé avec Laurence Hibrant, Koji Kawamura et toute l'équipe de génétique et horticulture de l'I.N.R.A. En utilisant l'analyse en composante principale à noyaux, nous avons détecté un nouveau QTL qui explique la variabilité d'architecture d'inflorescence d'une population de rosiers Rosa Wichurana. Actuellement, en collaboration avec "Les amies de la Roseraie du Val-de-Marne", nous procédons à un vrai travail de taxinomie d'une classe de rosiers appelés rosiers Noisettes, découvert par Louis Claude Noisette en 1814, grâce à un tableau de données de 67 variables décrivant les caractères phénotypiques de ces rosiers. Plus récemment, j'ai aussi pris contact avec le pôle santé de l'université et le monde industriel. Ces collaborations montrent toute la diversité des applications potentielles des statistiques. La fin du dernier chapitre aborde ces problématiques de manière synthétique.

La synthèse de ces différents travaux est précédée d'un chapitre introductif en français présentant le cadre mathématique. Ce chapitre confronte notamment les deux paradigmes de l'apprentissage : statistique ou en ligne. Ce chapitre est aussi rédigé pour tout mathématicien curieux qui, au fond, n'a pas une idée très claire de ce qu'est, et ce que devient les statistiques depuis le début du XXème siècle. En un mot, les statistiques tentent depuis plus d'un siècle de s'éloigner des modèles trop restictifs, en considérant successivement des modèles gaussiens, puis paramétriques, puis non-paramétriques, jusqu'à récemment des modèles sans aucune hypothèse probabiliste. Ainsi, dans cette courte introduction, quelques résultats fondamentaux (inégalités de Vapnik, bornes de regrets) sont énoncés, au sujet de quelques méthodes séminales (minimiseur du risque empirique, mélange à poids exponentiels).

Les résultats majeurs de ce travail seront accompagnés de schémas de preuve, où les aspects techniques seront mis de côté pour alléger la rédaction. Les preuves complètes sont disponibles en ligne dans les articles en questions (à chaque section est associé un ou plusieurs articles).

Enfin, les dernières pages compilent les problèmes ouverts auxquels j'ai pensé lors de l'écriture de ce mémoire. La rédaction est un travail parfois fastidieux mais qui permet de prendre du recul et de faire le point sur l'état de la recherche. Ainsi, comme le montre la quinzaine de problèmes ouverts, il y a encore du pain sur la planche pour les prochaines années !

Listes des travaux

Articles publiés ou acceptés

[L1] S. Loustau. Aggregation of SVM classifiers using Sobolev spaces. *Journal of Machine Learning Research.* 9 : 1559-1982, 2008.

[L2] S. Loustau. Penalized empirical risk minimization over Besov spaces. *Electronic Journal of Stats.* 3 : 824-850, 2009.

[L3] S. Loustau and C. Marteau. Minimax fast rates in discriminant analysis with errors in variables. *Bernoulli.* To appear.

[L4] S. Loustau. Inverse Statistical Learning. *Electronic Journal of Stats* 7 : 2065-2097, 2013.

[L5] K. Kawamura, L. Hibrand-Saint-Oyant, F. Foucher, T. Thouroude, and S. Loustau. Kernel methods for phenotyping complex plant architecture. *Journal of Theoretical Biology.* 342 : 83-92, 2014.

[L6] M. Chichignoud and S. Loustau. Adaptive noisy clustering. *IEEE Transaction on Information Theory.* 60 (11) : 1-14, 2014.

Articles soumis ou en révision

[L7] M. Chichignoud and S. Loustau. Bandwidth selection in kernel empirical risk minimization via the gradient. *Annals of Statistics, in revision.* 2014.

[L8] S. Loustau and C. Marteau. Noisy classification with boundary assumptions *Soumis.* 2013.

[L9] S. Loustau. Online clustering of individual sequences. *Soumis.* 2014.

[L10] C. Brunet and S. Loustau. Noisy quantization : theory and practice. *Soumis.* 2014.

[L11] S. Loustau. Minimax online bi-clustering. *Soumis.* 2014.

[L12] S. Loustau and S. Souchet. Two bandwidth selection methods for noisy clustering. *Soumis.* 2014.

[L13] P. and S. Loustau. In-game prediction with SVM. *Soumis.* 2014.

Autres travaux

[L14] S. Loustau. Performances statistiques de méthodes à noyaux. *Thèse de doctorat de l'Université de Provence.* 2008.

[L15] S. Loustau. Model selection in Kernel Projection Machines. *Preprint.* 2008.

[L16] S. Loustau. Anisotropic oracle inequalities in noisy quantization. *Preprint.* 2013.

[L17] H. Roman, M. Rapicault, M. Larenaudie, K. Kawamura, T. Thouroude, A. Chastellier, A. Lemarquand, F. Dupuis, F. Foucher, S. Loustau and L. Hibrand-Saint Oyant. Analyses and genetic determinism of floral characters in rose. *En cours.* 2014.

Table des matières

Chapitre 1

Introduction générale

Ce chapitre introduit le cadre mathématique de ce manuscrit, de l'apprentissage statistique à l'apprentissage en ligne. On verra dans cette introduction que des liens assez forts existent entre ces deux types d'apprentissage de prime abord bien distincts. Ces liens verront le jour à la lumière du lemme d'Hoeffding, du phénomène des vitesses rapides, ou encore de l'adaptation.

Le problème de l'adaptation découle de toute procédure statistique. Cela concerne la calibration des méthodes, qui possèdent - sauf exception - des paramètres à fixer. Cette problématique est aujourd'hui encore très populaire en statistique mathématique et en apprentissage. On présentera les travaux pionniers à ce sujet et l'application à des problèmes d'apprentissage.

Enfin, ce chapitre sera l'occasion d'introduire les principaux résultats de ce manuscrit, de l'apprentissage statistique de problème inverse à l'apprentissage en ligne, en passant par l'adaptation, le choix de la fenêtre et la considération de problèmes réels.

Contents

1.1 L'apprentissage statistique et l'apprentissage en ligne

Généralités

La théorie de l'apprentissage s'intéresse à la construction et à l'évaluation d'algorithmes d'aide à la décision basés sur une suite d'observations. Deux paradigmes dominent les travaux mathématiques sur le sujet. Ils concernent la collecte de cette suite d'observations :

— l'apprentissage statistique (*statistical learning*) considère un échantillon de n variables aléatoires généralement indépendantes et identiquement distribuées (i.i.d.). L'algorithme prend une décision (estimation, classification, test) à partir de cette suite d'observations [1]. On doit ses fondements mathématiques aux travaux de Vladimir Vapnik et Alexei Chervonenkis (VC theory, voir Vapnik and Chervonenkis [1971], et aussi Vapnik [2000]). On peut citer Devroye, Györfi, and Lugosi [1996] pour un ouvrage introductif sur le sujet. Depuis les prémices de la statistique mathématique introduite au début du XXème siècle, la tendance actuelle est de s'éloigner des modèles trop restrictifs.

— l'apprentissage en ligne (*online learning*) considère les observations arrivant de manière séquentielle, et très souvent sans aucune hypothèse probabiliste. L'algorithme est alors séquentiel et répond à chaque observation, à la manière d'un jeu contre la nature. Cette discipline est à l'intersection de la statistique, de l'informatique et de la théorie des jeux. Ses travaux fondateurs sont dus à

[1] On parle aussi du mode "batch", puisqu'on prend une décision à partir de ce paquet d'observations.

Hannan (Hannan [1957]) ou plus récemment Littlestone et Warmuth (Littlestone and Warmuth [1994]). Pour un survol des principaux résultats sur le sujet, on peut citer Cesa-Bianchi and Lugosi [2006]. Généralement, les algorithmes proposent un mélange d'avis d'experts mis à jour à chaque nouvelle observation.

Les observations sont le point de départ du statisticien. Comme on vient de le voir, deux points de vue bien différents sont proposés en apprentissage pour modéliser cette collecte d'informations. Il s'en suit deux manières distinctes de (1) répondre à la problématique d'aide à la décision et (2) mesurer les performances des algorithmes sous-jacents.

En apprentissage statistique, on considère un échantillon i.i.d. de variables aléatoires $\mathcal{D}_n := \{\mathcal{Z}_1, \ldots, \mathcal{Z}_n\}$ de loi inconnue P. Le risque d'une règle de décision f (estimateur, classifieur, test) est mesuré par une fonction de perte intégrée sous la loi P, c'est-à-dire par la quantité :

$$R(f) = \mathbb{E}_P \ell(f, \mathcal{Z}), \tag{1.1}$$

où $\ell(f, z)$ mesure la perte de f associée au point d'observation z et $\mathcal{Z}$ est une variable aléatoire de loi P indépendante de $\mathcal{D}_n$. Ce risque est aussi appelé erreur de généralisation. Étant en présence d'un phénomène stationnaire, c'est la perte moyenne qu'engendrera f si l'on observe une nouvelle variable aléatoire $\mathcal{Z}$ de loi P indépendante de $\mathcal{D}_n$ [2].

En apprentissage en ligne, le risque (1.1) n'est pas disponible puisqu'on ne suppose aucun modèle probabiliste sur la suite d'observations. A chaque tour $t = 1, \ldots, T$, où T est appelé l'horizon, on va proposer une décision. Dans le jeu de prédiction avec avis d'experts, à chaque tour t, on veut prédire z_t à partir des observations passées $z_1, \ldots, z_{t-1}$ et d'avis d'experts $p_{t,1}, \ldots, p_{t,N}$, où N est le nombre d'experts. A la fin du jeu, la performance de notre algorithme est mesuré par le regret :

$$\sum_{t=1}^{T} \ell(\hat{z}_t, z_t) - \min_{k=1,\ldots,N} \sum_{t=1}^{T} \ell(p_{t,k}, z_t), \tag{1.2}$$

où $\ell(z', z)$ est la perte de z' associée à l'observation z. Le regret mesure la différence entre la perte cumulée de l'algorithme et la perte cumulée du meilleur expert parmi les N experts qui donnent leur avis à chaque tour. Cette notion de perte relative est elle aussi présente en apprentissage statistique, où l'on considère habituellement l'excès de risque :

$$R(f) - R(f^\star), \tag{1.3}$$

où $f^\star$ est la meilleure règle de décision - appelée règle de Bayes - pour le problème considéré. L'introduction de cette quantité permet de tenir compte de la difficulté intrinsèque du problème, c'est-à-dire indépendamment de la méthode utilisée.

Deux résultats clés

Les résultats obtenus en apprentissage dépendent du type d'apprentissage considéré : statistique ou en ligne. En apprentissage statistique, on s'intéresse à contrôler l'excès de risque (1.3) alors qu'en apprentissage en ligne, on veut contrôler le regret (1.2). Ces deux problèmes sont sensiblement différents et aboutissent à des méthodes différentes.

Si l'on dispose d'un échantillon i.i.d. $\mathcal{D}_n = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_n\}$, et d'un ensemble d'hypothèses $\mathcal{F}$, un candidat naturel est le minimiseur du risque empirique (ERM) :

$$\widehat{f} := \arg\min_{f \in \mathcal{F}} \widehat{R}(f) \text{ avec } \widehat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f, \mathcal{Z}_i). \tag{1.4}$$

En effet, on peut montrer facilement que $\widehat{f}$ vérifie par définition :

$$R(\widehat{f}) - R(f^\star_{\mathcal{F}}) \le 2 \sup_{f \in \mathcal{F}} \left| R(f) - \widehat{R}(f) \right|, \tag{1.5}$$

[2] Au contraire, l'erreur d'apprentissage est l'erreur calculée sur l'échantillon $\mathcal{D}_n$, en remplaçant la mesure P par la mesure empirique $P_n = 1/n \sum \delta_{\mathcal{Z}_i}$ dans (1.1). Minimiser l'erreur d'apprentissage peut conduire au phénomène de sur-apprentissage.

où $f^\star_{\mathcal{F}}$ est le minimiseur du risque $R(\cdot)$ sur $\mathcal{F}$. Cette inégalité motive l'introduction des processus empiriques et l'étude de lois des grands nombres uniformes. Le résultat suivant assure une inégalité en grande probabilité pour l'excès de risque de l'estimateur (1.4) dans le cadre de la classification binaire.

Théorème 1 (Inégalité de Vapnik (1971))**.** *Soit $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ un échantillon i.i.d. de loi P sur $\mathbb{R}^d \times \{0,1\}$. Soit $\mathcal{F}$ un ensemble de classifieurs de la forme $f : \mathbb{R}^d \to \{0,1\}$. Alors, on a :*

$$\mathbb{P}_{\mathcal{D}_n}\left(\sup_{f\in\mathcal{F}}\left|R(f) - \widehat{R}(f)\right| > \epsilon\right) \leq 8\mathcal{S}(\mathcal{F}, n)e^{-\frac{n\epsilon^2}{32}},$$

où $\mathcal{S}(\mathcal{F}, n)$ est le nième coefficient d'éclatement de $\mathcal{F}$. De plus, si la dimension de Vapnik de $\mathcal{F}$, notée $\mathrm{VC}(\mathcal{F})$ *est finie, on a :*

$$\mathbb{E}_{\mathcal{D}_n} R(\widehat{f}) - R(f^\star_{\mathcal{F}}) \leq 16\sqrt{\frac{4 + \mathrm{VC}(\mathcal{F})\log n}{2n}}.$$

Cette version de l'inégalité de Vapnik est exposée dans Devroye, Györfi, and Lugosi [1996]. L'inégalité remonte à Vapnik and Chervonenkis [1971].

Si l'on considère un jeu séquentiel où à chaque tour $t \in \mathbb{N}^*$, nous disposons des observations passées $z_1, \ldots, z_{t-1}$ et d'avis d'experts $p_{t,1}, \ldots, p_{t,N}$, l'approche la plus répandue est de mélanger les avis des N experts ou de suivre le vote majoritaire. Littlestone and Warmuth [1994] s'intéressent à ces deux alternatives et introduisent par exemple, à chaque tour $t = 1, \ldots, T$, la prédiction :

$$\hat{z}_t = \sum_{k=1}^{N} w_{k,t-1}p_{t,k} \text{ avec } w_{k,t-1} = \frac{e^{-\lambda\sum_{u=1}^{t-1}\ell(p_{u,k},z_u)}}{W_{t-1}}, \tag{1.6}$$

où W_{t-1} est tel que $\sum w_{k,t-1} = 1$ et $\lambda > 0$ est un paramètre de température inverse. On peut montrer le résultat suivant :

Théorème 2 (Borne de regret)**.** *Soit $\mathcal{Y} \subseteq [0,1]$ et $\ell(\cdot, z)$ convexe quel que soit $z \in \mathcal{Y}$. Alors la prévision à poids exponentiels* (1.6) *vérifie :*

$$\sum_{t=1}^{T}\ell(\hat{z}_t, z_t) - \min_{k=1,\ldots,N}\sum_{t=1}^{T}\ell(p_{t,k}, y_t) \leq \frac{\log N}{\lambda} + \frac{\lambda T}{8} = \sqrt{\frac{T\log N}{2}}, \tag{1.7}$$

où la dernière égalité a lieu en prenant $\lambda^* = \sqrt{(8\log N)/T}$.

Ce résultat est présenté dans Cesa-Bianchi and Lugosi [2006]. Il est intéressant de comparer le Théorème 1 au Théorème 2. A première vue, ces résultats sont sensiblement différents : l'un est probabiliste, et a lieu avec grande probabilité (ou en espérance), par rapport à la loi de l'échantillon. L'autre est entièrement déterministe, et a lieu pour toute suite d'observations. Cela dit, ces deux résultats sont basés sur le même outil probabiliste : le lemme de Hoeffding.

Lemme 1. *Soit X une variable aléatoire réelle telle qu'il existe deux réels a, b tels que $a \leq X \leq b$ p.s. Alors :*

$$\log \mathbb{E}e^{\lambda X} \leq \lambda\mathbb{E}X + \frac{\lambda^2(b-a)^2}{8}.$$

Ce lemme est à l'origine de l'inégalité de Hoeffding, qui concerne la concentration d'une somme de variables aléatoires i.i.d. vers son espérance (inégalité de concentration). En appliquant cette inégalité à $\widehat{R}(f)$, on peut montrer l'inégalité de Vapnik en utilisant (1.5) et l'hypothèse $\mathrm{VC}(\mathcal{F}) < \infty$.

De manière plus surprenante, bien que la suite (z_t) soit déterministe, la preuve de l'inégalité (1.7) utilise le Lemme 1 de la manière suivante. On peut noter que (1.6) est un mélange convexe de la suite $\mathbf{p}_t = (p_{t,k})_{k=1}^N$. Ainsi, on peut écrire, en utilisant le lemme pour X de loi discrète $(w_{t,k})_{k=1}^N$ à valeurs dans $(p_{t,k})_{k=1}^N$:

$$\frac{1}{\lambda}\log\frac{W_t}{W_{t-1}} = \frac{1}{\lambda}\log\mathbb{E}_{\mathbf{w}_t}e^{-\lambda\ell(\mathbf{p}_t,z_t)} \leq -\mathbb{E}_{\mathbf{w}_t}\ell(\mathbf{p}_t, z_t) + \frac{\lambda}{8} \leq -\ell(\hat{z}_t, z_t) + \frac{\lambda}{8},$$

où l'on a utilisé la convexité de $\hat{z} \mapsto \ell(\hat{z}, z)$ pour obtenir la dernière inégalité. Cette série d'inégalités est à l'origine de la borne du Théorème 2 (Cesa-Bianchi and Lugosi [2006] pour une preuve complète).

En première conclusion, bien qu'a priori très distincts, apprentissage statistique et apprentissage en ligne utilisent les mêmes fondements probabilistes : la concentration de la mesure. L'inégalité (1.5) motive la théorie des processus empiriques, qui est au centre des bornes d'excès de risque en apprentissage statistique. En apprentissage en ligne, on peut considérer que la prise de décision comporte un aléa, puisqu'on utilise des mélanges d'experts (ou plus généralement dans la suite des estimateurs randomisés).

1.2 Localisation, marge et vitesses rapides

La section précédente a présenté deux résultats fondateurs de la théorie de l'apprentissage. L'inégalité de Vapnik induit une convergence de l'excès de risque vers 0 à vitesse $\log n/\sqrt{n}$ alors que le Théorème 2 entraîne une convergence de la perte moyenne de l'algorithme (1.6) vers la perte moyenne du meilleur expert à vitesse $1/\sqrt{T}$ (en normalisant l'inégalité par le nombre de tours T). Dans cette section, on va voir que des vitesses de convergence plus rapides sont possibles sous certaines hypothèses.

Une procédure de localisation

Dans le Chapitre 2 de ce manuscrit, on veut établir des bornes pour l'excès de risque (1.3), où $\mathcal{F}$ est un ensemble très souvent fonctionnel (de dimension infinie) et VC($\mathcal{F}$) est remplacée par des hypothèses sur l'entropie de $\mathcal{F}$. Dans ce cadre, on utilisera un raffinement de l'inégalité de Vapnik, en réduisant le supremum dans l'inégalité (1.5) aux fonctions $f \in \mathcal{F}(\delta) := \{f \in \mathcal{F} : R(f) - R(f^\star_{\mathcal{F}}) \leq \delta\}$ pour $\delta > 0$. En effet, on peut remarquer que si $\ell(f, z)$ est à valeurs dans $[0, \delta_0]$:

$$\begin{aligned} R(\widehat{f}) - R(f^\star_{\mathcal{F}}) &\leq \sup_{f \in \mathcal{F}(\delta_0)} \left|(R - \widehat{R})(f - f^\star_{\mathcal{F}})\right| := \Psi_n(\delta_0) \\ &\leq \mathbb{E}\Psi_n(\delta_0) + \mathcal{U}_n(\delta_0, t), \end{aligned}$$

où la dernière inégalité a lieu avec probabilité $1 - e^{-t}$ grâce à une inégalité de concentration de Talagrand (Talagrand [1995]). Le terme $\mathcal{U}_n(\delta, t)$ s'écrit par exemple en utilisant la version de Bousquet (Bousquet [2002]) :

$$\mathcal{U}_n(\delta, t) = \sqrt{\frac{2t}{n}\left[\sigma^2(\mathcal{F}) + (1 + \delta_0)\mathbb{E}\Psi_n(\delta)\right]} + \frac{t}{3n},$$

où $\sigma^2(\mathcal{F}) = \sup_{\mathcal{F}} \mathrm{Var}_P\left[\ell(f, \mathcal{Z}) - \ell(f^\star_{\mathcal{F}}, \mathcal{Z})\right]$. On peut alors poser $\delta_1 = \mathbb{E}\Psi_n(\delta_0) + \mathcal{U}_n(\delta_0, t)$ et répéter la manœuvre pour obtenir après N itérations la borne suivante, avec probabilité $1 - Ne^{-t}$:

$$R(\widehat{f}) - R(f^\star_{\mathcal{F}}) \leq \mathbb{E}\Psi_n(\delta_{N-1}) + \mathcal{U}_n(\delta_{N-1}, t) = \delta_N.$$

Avec un choix de N assez grand, la détermination d'un contrôle optimal du membre de gauche revient à résoudre l'équation du point fixe $\psi_n(\delta) = \delta$ où $\psi_n(\delta) = \mathbb{E}\Psi_n(\delta) + \mathcal{U}_n(\delta, t)$.

Cette heuristique est à l'origine des vitesses de convergence rapides, c'est-à-dire plus rapide que $1/\sqrt{n}$ en apprentissage statistique. En effet, en résolvant l'équation ci-dessus, on peut obtenir avec grande probabilité :

$$R(\widehat{f}) - R(f^\star_{\mathcal{F}}) \lesssim n^{-\kappa/(2\kappa+\rho-1)}, \tag{1.8}$$

où $\rho > 0$ mesure la complexité de $\mathcal{F}$ en terme de vitesse d'entropie[3] et où l'on suppose l'existence d'un paramètre de marge $\kappa \geq 1$ tel que :

$$\mathrm{Var}_P\left[\ell(f, \mathcal{Z}) - \ell(f^\star_{\mathcal{F}}, \mathcal{Z})\right] \lesssim \left[R(f) - R(f^\star_{\mathcal{F}})\right]^{1/\kappa}. \tag{1.9}$$

L'inégalité ci-dessus est centrale dans la procédure de localisation pour majorer la variance $\sigma^2(\mathcal{F})$ qui apparaît dans l'inégalité de concentration. Cela permet d'obtenir une borne supérieure de $\mathcal{U}_n(\delta, t)$ et

[3] Le paramètre ρ permet de contrôler la complexité de la classe $\mathcal{F}$ de manière à majorer le terme $\mathbb{E}\Psi_n(\delta)$ grâce à des techniques de chaînage.

finalement des vitesses de convergence rapides. De nombreux travaux ont établi des vitesses de convergence de ce type en apprentissage sous des hypothèses de marge. En classification binaire, on peut citer les travaux précurseurs de Mammen and Tsybakov [1999], Tsybakov [2004], Tsybakov and van de Geer [2005], mais aussi Massart and Nédélec [2006], Blanchard, Bousquet, and Massart [2008], Blanchard, Lugosi, and Vayatis [2003] ou encore Bartlett and Mendelson [2006] ou Bartlett, Bousquet, and Mendelson [2005]. Les travaux les plus proches de la méthode de localisation présentée ici remontent à Koltchinskii and Panchenko [2000] (voir aussi Koltchinskii [2006]).

Le gradient comme alternative

Dans ce mémoire, on présente un autre moyen d'obtenir des vitesses rapides pour l'excès de risque lorsque dim $\mathcal{F} < \infty$. Dans ce cas, on introduit un nouveau critère de performances pour le minimiseur du risque empirique (1.4) appelé le gradient de l'excès de risque. Cette quantité est définie dans le Chapitre 3 par :

$$|G(f_\theta) - G(f_{\theta^\star})|_2 = |\nabla R(f_\theta) - \nabla R(f_{\theta^\star})|_2,$$

où $\nabla R(f_\theta)$ est le gradient de l'application $\theta \mapsto R(f_\theta)$. La notation f_θ signifie que chaque élément de $\mathcal{F}$ est une fonction qui ne dépend que d'un nombre fini de paramètre $\theta \in \mathbb{R}^m$. L'introduction du gradient du risque va nous permettre d'obtenir un contrôle de l'excès de risque grâce à l'inégalité suivante (Lemme 5 du Chapitre 3) :

$$\sqrt{R(f_\theta) - R(f_{\theta^\star})} \lesssim \lambda_{\min}^{-1} |G(f_\theta) - G(f_{\theta^\star})|_2, \tag{1.10}$$

où $\lambda_{\min}$ est la plus petite valeur propre de la Hésienne $\mathcal{H}_R$ du risque, où $\theta \mapsto R(f_\theta)$ est supposé de classe $\mathcal{C}^2(U)$ et $U \subset \mathbb{R}^m$ est un voisinage de $\theta^\star$. Cette inégalité, basée sur la régularité du risque, permet d'obtenir cette série d'inégalités :

$$\sqrt{R(f_{\widehat{\theta}}) - R(f_{\theta^\star})} \lesssim |G(f_{\widehat{\theta}}) - G(f_{\theta^\star})|_2 \leq \sup_{\theta \in U} \left| \widehat{G}(f_\theta) - G(f_\theta) \right| \lesssim n^{-1/2}, \tag{1.11}$$

où $f_{\widehat{\theta}}$ est défini en (1.4) et supposé proche de $f_{\theta^\star}$. Ainsi, on remarque qu'en utilisant (1.11), l'obtention de vitesses rapides pour l'excès de risque est assurée grâce à une simple inégalité de Vapnik du type (1.5). La machinerie de la localisation et l'hypothèse de marge (1.9) ne sont donc plus nécessaires à l'obtention de vitesses de convergence rapides.

Le cas de l'apprentissage en ligne

L'obtention de vitesses rapides en apprentissage en ligne est antérieure aux travaux de Mammen and Tsybakov [1999] en classification. En effet, dès les travaux de Vovk [2001], des bornes de regret indépendantes de T ont été obtenues pour la perte des moindres carrés. Des résultats similaires sont obtenus sous des hypothèses très fortes sur la perte dans Haussler, Kivinen, and Warmuth [1998]. Ces hypothèses apparaissent sous plusieurs formes dans la littérature, et l'on peut noter des similitudes avec le cadre statistique précédent. En effet, du point de vue minimax, nous savons que les vitesses optimales en apprentissage en ligne dépendent de la régularité de la fonction de perte. Avec des hypothèses faibles sur la régularité de la perte, Haussler, Kivinen, and Warmuth [1998] montrent de manière générale que :

$$(1+o(1))c_L\sqrt{T\log N} \leq \inf_{\hat{z}_t} \sup_{z_t,\mathbf{p}_t} \sum_{t=1}^T \left\{ \ell(\hat{z}_t, z_t) - \min_{k=1,\dots,N} \sum_{t=1}^T \ell(p_{t,k}, z_t) \right\} \leq c_L\sqrt{T\log N}.$$

Ces bornes avaient déjà été démontrées par Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth [1997] pour la perte logarithmique. Les résultats de Haussler, Kivinen, and Warmuth [1998] sont plus généraux, et proposent aussi un regret indépendant de T sous des conditions spécifiques (dont une perte deux fois différentiable) de la forme :

$$(1+o(1))c_L\sqrt{\log N} \leq \inf_{\hat{z}_t} \sup_{z_t} \sum_{t=1}^T \left\{ \ell(\hat{z}_t, z_t) - \min_{k=1,\dots,N} \sum_{t=1}^T \ell(p_{t,k}, z_t) \right\} \leq c_L\sqrt{\log N}.$$

Un regard attentif sur la preuve de la première inégalité (appelée borne inférieure) nous montre que pour obtenir un résultat de cette nature, Haussler, Kivinen, and Warmuth [1998] supposent l'existence d'un unique minimiseur du risque, où le supremum ci-dessus est minoré par une espérance (voir le Chapitre 4 pour ces aspects minimax). Cette hypothèse est très proche d'une Hessienne définie-positive dans l'approche du gradient. Des bornes similaires du regret (indépendantes de T) sont démontrées dans Cesa-Bianchi and Lugosi [2006] lorsque la fonction $\hat{z} \mapsto e^{-\lambda \ell(\hat{z},z)}$ est concave, ce qui reste bien une hypothèse de régularité sur la perte, et notamment sur le signe de sa dérivée seconde. Dans Audibert [2009], une hypothèse de variance est utilisée pour obtenir des résultats similaires (Théorème 4 ci-dessous).

Plus récemment, Rakhlin, Sridharan, and Tewari [2010] s'intéressent aux procédures de localisation décrites au début de la section dans un contexte d'apprentissage en ligne. En effet, dans le cadre de la prévision de suites individuelles, on peut utiliser des techniques de localisation pour obtenir des vitesses de convergence rapides. Dans ce cas, la complexité étudiée dans Rakhlin, Sridharan, and Tewari [2010] est une version séquentielle de $\mathbb{E}\Psi_n(\delta)$ (complexité de Rademacher locale séquentielle).

1.3 La théorie PAC-Bayésienne

L'approche PAC-Bayésienne garantit les performances de règles de décisions randomisées. Les résultats ont lieu sans aucune hypothèse sur la suite d'observations $\mathcal{D}_n = \{\mathcal{Z}_1, \dots, \mathcal{Z}_n\}$, outre l'hypothèse i.i.d. de l'apprentissage statistique. L'approche PAC-Bayésienne diffère en ce sens de l'approche Bayésienne classique où les résultats sont obtenus sous l'hypothèse d'une loi a priori.

L'inégalité de Mac Allester

Les fondements de l'approche PAC-Bayésienne sont dus aux travaux de Mac Allester (voir Mac Allester [1998]). Le principe de l'approche PAC-Bayésienne est de construire une règle de décision randomisée $f \sim \rho$, où $\rho := \rho(\mathcal{Z}_1, \dots, \mathcal{Z}_n)$ est une mesure aléatoire définie sur l'ensemble $\mathcal{F}$. On veut établir des résultats du type :

$$\forall \epsilon > 0,\ \forall \pi \in \mathcal{M}_1^+(\mathcal{F}),\ \mathbb{P}_{\mathcal{D}_n}(\mathbb{E}_{f\sim\rho} R(f) \leq \mathcal{B}(\rho,\pi)) \geq 1-\epsilon. \tag{1.12}$$

Contrairement à l'inégalité de Vapnik qui utilise la théorie des processus empiriques, dans la théorie PAC-Bayésienne, un rôle majeur est joué par la divergence de Kullback $\mathcal{K}(\rho,\pi)$. En effet, la relation de dualité convexe suivante est l'argument majeur de la théorie PAC-Bayésienne [4] (on peut citer Rockafellar [1970] pour les fondements issus de l'analyse convexe) :

$$\log \mathbb{E}_{f\sim\pi} e^{h(f)} = \sup_{\rho \in \mathcal{M}_1^+(\mathcal{F})} \left\{ \mathbb{E}_{f\sim\rho} h(f) - \mathcal{K}(\rho,\pi) \right\}.$$

Par exemple, l'égalité précédente est le principal ingrédient du théorème suivant, souvent présenté comme un résultat pionnier de la théorie PAC-Bayésienne.

Théorème 3 (Mac Allester [1998])**.** *Soit $\mathcal{D}_n = \{Z_1, \dots, \mathcal{Z}_n\}$ un échantillon i.i.d. de loi P, $\mathcal{F}$ un ensemble de règles de décision et $\ell(f,\cdot)$ à valeurs dans* $[0,1]$*. Avec les notations de la section précédente, quel que soit $\epsilon > 0$, pour tout prior $\pi \in \mathcal{M}_1^+(\mathcal{F})$, avec probabilité (sur l'échantillon $\mathcal{D}_n$) au moins $1-\epsilon$, quel que soit $\rho \in \mathcal{M}_1^+(\mathcal{F})$:*

$$\mathbb{E}_{f\sim\rho} R(f) \leq \mathbb{E}_{f\sim\rho} \widehat{R}(f) + \sqrt{\frac{\log(4n\epsilon^{-1}) + \mathcal{K}(\rho,\pi)}{2n-1}}.$$

Ce théorème est intéressant ici pour deux raisons. Tout d'abord, sa preuve fait appel à des outils alternatifs à la théorie des processus empiriques et aux travaux de Vapnik et Chervonenkis. Ces outils constituent le fondement théorique des résultats d'apprentissage en ligne, et notamment de la prédiction de suites individuelles. On pourra se référer à Seeger [2008] pour une preuve élégante utilisant successivement l'inégalité de Markov, la formule de dualité ci-dessus et le Lemme 1, comme dans la preuve du Théorème 2 ! Enfin, ce résultat motive l'utilisation de poids exponentiels du type (1.6) en apprentissage.

[4] Dans la formule de dualité, $\mathbb{E}_{f\sim\rho} h(f) = \sup_{A\in\mathbb{R}} \min(A, h(f))$, et par convention le membre de droite vaut $-\infty$ lorsque $\mathcal{K}(\rho,\pi) = +\infty$ (voir par exemple Catoni [2001].)

En effet, le Théorème 3 a lieu pour toute mesure ρ. Il est donc naturel de chercher l'expression de ρ qui minimise le membre de droite de la borne ci-dessus. La mesure ρ qui réalise ce minimum n'est autre qu'une mesure de Gibbs de la forme :

$$d\widehat{\rho}(f) := \frac{e^{-\lambda \widehat{R}(f)}}{W_\pi} d\pi(f), \tag{1.13}$$

où W_π est la constante de normalisation et $\lambda > 0$ est un paramètre de température (inverse).

Quelques résultats récents de la théorie PAC-Bayésienne

Depuis les travaux de Mac-Allester, de nombreux auteurs ont obtenu des résultats PAC-Bayésiens en apprentissage statistique. L'ouvrage d'Olivier Catoni (Catoni [2001]) propose des raffinements de cette inégalité, en utilisant l'inégalité de Bernstein par exemple. Plus récemment, Audibert [2009] a montré une inégalité de la même forme pour un algorithme de type moyenne miroir (mirror averaging). On présente cette approche dans la suite de cette section qui sera utilisée dans le Chapitre 4 en apprentissage en ligne. En effet, ces travaux permettent de traiter de manière unifiée les bornes d'excès de risque en apprentissage statistique (avec parfois des résultats plus fins) et les bornes de regrets en apprentissage en ligne. On peut déjà conclure que c'est bien la théorie PAC-Bayésienne, inspirée des travaux de Mac-Allester en apprentissage statistique, qui constitue les fondements théoriques de la prévision de suites individuelles.

On considère une suite déterministe $z_1, \ldots, z_T \in \mathbb{R}$ où T est l'horizon connu du statisticien, un ensemble de règles de décision $\mathcal{F}$ et une fonction de perte $\ell(f, z)$ qui mesure la perte de f au point d'observation z. Dans ce cadre en ligne et déterministe, la théorie PAC-Bayésienne de Audibert [2009] propose à chaque tour t de construire une fonction $\widehat{f}_t \sim \widehat{\rho}_t$ où $\widehat{\rho}_t := \widehat{\rho}_t(z_1, \ldots, z_{t-1}, \widehat{f}_1, \ldots, \widehat{f}_{t-1})$ est une mesure telle que $\mathbb{E}_{(\widehat{\rho}_1,\ldots,\widehat{\rho}_t)} \ell(\widehat{f}_t, z_t)$ soit petite. Plus précisément, l'algorithme ainsi construit satisfait la borne de regret suivante :

Théorème 4 (Audibert [2009])**.** *Soit $z_1, \ldots, z_T$ une suite déterministe et $\mathcal{F}$ un ensemble de règles de décision. Alors, si la fonction ℓ est λ-exponentielle concave, on a :*

$$\sum_{t=1}^{T} \mathbb{E}_{(\widehat{\rho}_1,\ldots,\widehat{\rho}_t)} \ell(\widehat{f}_t, z_t) \leq \inf_{\rho \in \mathcal{M}_1^+(\mathcal{F})} \left\{ \mathbb{E}_{\theta\sim\rho} \sum_{t=1}^{T} \ell(f, z_t) + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\}.$$

De plus, sans aucune hypothèse sur la perte, quel que soit $\lambda > 0$:

$$\sum_{t=1}^{T} \mathbb{E}_{(\widehat{\rho}_1,\ldots,\widehat{\rho}_t)} \ell(\widehat{f}_t, z_t) \leq \inf_{\rho \in \mathcal{M}_1^+(\mathcal{F})} \left\{ \mathbb{E}_{\theta\sim\rho} \sum_{t=1}^{T} \ell(f, z_t) + \frac{\lambda}{2} \mathbb{E}_{f\sim\rho} \sum_{t=1}^{T} \mathbb{E}_{(\widehat{\rho}_1,\ldots,\widehat{\rho}_t)} \left(\ell(f, z_t) - \ell(\widehat{f}_t, z_t) \right)^2 + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\}.$$

Les deux bornes de regrets ci-dessus sont la version déterministe des inégalités PAC-Bayésiennes de la forme (1.12). Le risque est remplacé par la perte cumulée et l'infimum à droite est dû au choix (1.13) de la mesure $\widehat{\rho}_t$ à chaque tour du jeu séquentiel. Ces deux bornes aboutissent à des vitesses de convergence différentes, comme discuté dans le paragraphe précédent. Dans le premier cas le plus favorable, la perte utilisée satisfait de bonnes propriétés et on obtient dans ce cas des vitesses de convergence rapides. Dans le deuxième cas, qui a lieu sans aucune hypothèse sur la fonction de perte, un terme supplémentaire apparaît dans l'inégalité PAC-Bayésienne et des vitesses de convergence lentes en découlent. On peut retrouver des inégalités probabilistes du cadre i.i.d. en ajoutant une étape de moyennisation supplémentaire. Cela permet d'utiliser la *chain rule* (Barron [1987]) comme dans le cadre séquentiel et la preuve du Théorème 2.

1.4 Adaptation, sélection de fenêtres

La théorie de l'apprentissage propose des algorithmes pour résoudre de nombreux problèmes. Malheureusement, ces outils d'aide à la décision possèdent des paramètres à fixer qui dictent leurs performances. Dans la section 1.1 de ce chapitre, nous avons introduit deux méthodes d'apprentissage : le minimiseur du risque empirique et le mélange d'experts à poids exponentiels. Ce dernier exhibe d'emblée un paramètre de température que doit fixer le statisticien. Le théoricien peut proposer grâce au Théorème 2 $\lambda^* = \sqrt{8(\log N)/T}$. Malheureusement, ce choix dépend de l'horizon T du problème séquentiel. Dans un contexte purement en ligne, cette quantité (la fin du jeu) n'est pas connue et se pose le problème de calibration pratique de ce paramètre. De plus, comme souvent, ce choix théorique (en supposant que T est connu) n'est pas satisfaisant. Puisque le théorème s'intéresse au pire des cas, le choix de température qui en découle est souvent trop pessimiste en pratique. En présence d'un échantillon i.i.d. d'observations, le problème d'adaptation est de nature différente, puisqu'on connaît à l'avance le nombre d'observations. Malheureusement, dans ce cas, les paramètres à calibrer dépendent du comportement de la loi P qui génère l' échantillon, et d'hypothèses invérifiables en pratique (régularité de la densité des observations, hypothèse de marge, niveau de bruit, etc). Un problème séminal en statistique mathématique est le problème du choix de la fenêtre.

La règle de Lepski

Le choix de la fenêtre est un problème très populaire en statistique mathématique. Depuis l'introduction des estimateurs à noyaux (Rosenblatt [1956] et Parzen [1962]), on cherche à calibrer de manière automatique une fenêtre $h \in \mathbb{R}$ (cas uni varié), ou une fenêtre $\vec{h} \in \mathbb{R}_+^d$ (cas multivarié). En estimation de densité, on considère :

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_h(X_i - x), \tag{1.14}$$

où $\mathcal{K}_h(\cdot) = 1/h\mathcal{K}(\cdot/h)$ est une h-dilatation d'un noyau $\mathcal{K}(\cdot)$. Pour construire une fenêtre $h \in \mathbb{R}$ à partir des observations, Lepski [1990] introduit dans un modèle de bruit blanc une méthode basée sur la comparaison d'estimateurs construits sur une grille de fenêtre. Cette méthode permet d'obtenir une fenêtre adaptative, c'est-à-dire qui ne dépend pas de la régularité de la fonction à estimer, et qui possède de bonnes propriétés théoriques. Dans Lepski [1991], une procédure générale est construite, s'appliquant par exemple aux estimateurs à noyaux d'une densité (Lepski [1992a], Lepski [1992b]). En considérant une famille $\{\widehat{f}_h, h \in \mathcal{H}\}$, où $\mathcal{H} \subseteq \mathbb{R}$ est une grille unidimensionnelle de fenêtre, le point de départ est une décomposition biais-variance de la forme :

$$\mathbb{E}|\widehat{f}_h - f| \leq v(h) + b(h), \tag{1.15}$$

où $|\cdot|$ est une semi-norme. Généralement, la fonction $v(\cdot)$ est une fonction décroissante de h et connue explicitement, alors que $b(\cdot)$ est une fonction croissante de h qui dépend d'un paramètre inconnu (comme l'indice de régularité de la fonction f à estimer). On peut alors considérer, puisque $v(h)$ est connue :

$$\widehat{h} = \max\{h \in \mathcal{H} : \ \forall h' \leq h, |\widehat{f}_{h'} - \widehat{f}_h| \leq Cv(h')\}, \tag{1.16}$$

où $C > 0$ est une constante à calibrer. Cette règle est motivée par l'heuristique suivante ($h' \leq h$) :

$$|\widehat{f}_{h'} - \widehat{f}_h| \sim |\widehat{f}_{h'} - f| + |f - \widehat{f}_h| \sim v(h') + b(h') + v(h) + b(h) \sim v(h') + b(h).$$

Ainsi, d'un point de vue asymptotique, la règle (1.16) sélectionne :

$$\text{le plus grand } h > 0 \text{ tel que } \forall h' \leq h, \ b(h) \leq (C-1)v(h'),$$

c'est-à-dire une estimation de $h^\star$ qui réalise le compromis biais-variance dans la décomposition (1.15).

De nombreux travaux ont utilisé cette règle pour obtenir des résultats adaptatifs en statistique mathématique. Plus récemment, quelques auteurs se sont inspiré de Lepski pour résoudre des problèmes d'adaptation en apprentissage statistique (Katkovnik and Spokoiny [2008], Tsybakov [2004], Koltchinskii [2006]), Brunel [2013], Dattner, Reiß, and Trabs [2013]). C'est l'objet de la première partie du Chapitre 3. L'application de cette règle en apprentissage en ligne reste un problème ouvert. La décomposition du regret dans le Théorème 2 suggère pourtant l'utilisation d'une version séquentielle de cette règle en apprentissage en ligne.

La règle de Goldenshluger et Lepski

La règle (1.16) est restreinte à un choix de fenêtre $h \in \mathbb{R}$, c'est-à-dire à des problèmes isotropes. En effet, dans un cadre multivarié, on peut montrer que si l'on suppose une régularité isotrope (identique dans chaque direction) de la densité des observations, le choix optimal théorique de la fenêtre dans la version multivariée de (1.14) est de la forme $(h_1, \ldots, h_d) = (h, \ldots, h)$ et la règle de Lepski s'applique. Par contre, ce n'est pas le cas lorsque la régularité de la densité dépend de la direction. Récemment, Goldenshluger et Lepski, dans une série d'articles (Goldenshluger and Lepski [2008], Goldenshluger and Lepski [2009] pour le bruit blanc gaussien et Goldenshluger and Lepski [2011] pour l'estimation d'une densité) proposent une règle générale pour sélectionner une fenêtre $\vec{h} \in \mathbb{R}_+^d$ (on écrit $h \in \mathbb{R}_+^d$ dans la suite). Etant donné une famille d'estimateurs $\{\widehat{f}_h(\cdot), h \in \mathcal{H} \subset \mathbb{R}_+^d\}$ vérifiant (1.14), Goldenshluger and Lepski [2011] introduisent un estimateur auxiliaire :

$$\widehat{f}_{h,h'}(x) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathcal{K}_h * \mathcal{K}_{h'}\right)(\mathcal{Z}_i - z), \tag{1.17}$$

où $*$ est le produit de convolution dans $\mathbb{R}^d$. La règle générale de sélection de fenêtre s'écrit :

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} \left\{ \widehat{b}(h) + \widehat{\delta}(h) \right\}, \tag{1.18}$$

où $\widehat{b}(h)$ et $\widehat{\delta}(h)$ sont des estimateurs du biais et de la variance dans la décomposition (1.15) construite à l'aide de l'estimateur auxiliaire (1.17). En effet, dans (1.18) :

$$\widehat{b}(h) := \sup_{h' \in \mathcal{H}} \left\{ |\widehat{f}_{h,h'} - \widehat{f}_{h'}| - \mathrm{maj}(h,h') \right\} \text{ et } \widehat{\delta}(h) = \sup_{h' \in \mathcal{H}} \mathrm{maj}(h', h).$$

Dans l'équation ci-dessus, un rôle majeur est joué par la fonction $(h,h') \mapsto \mathrm{maj}(h,h')$ appelée majorant. Ce terme majore uniformément et avec grande probabilité la somme $|\widehat{f}_{h,h'} - \mathbb{E}_{\mathcal{D}_n}\widehat{f}_{h,h'}| + |\widehat{f}_h - \mathbb{E}_{\mathcal{D}_n}\widehat{f}_h|$ de sorte que :

$$|\widehat{f}_{h,h'} - \widehat{f}_{h'}| - \mathrm{maj}(h,h') \sim |f_{h,h'} - f_{h'}|.$$

Ainsi, en prenant le supremum sur $h' \in \mathcal{H}$, $\hat{b}(\cdot)$ est bien une estimation du biais en soulignant que :

$$\sup_{h'} \left| f_{h,h'} - f_{h'} \right| = \sup_{h'} |\mathcal{K}_{h'} * (f_h - f)| = b(h).$$

Dans cette dernière série d'égalités, une condition nécessaire est la linéarité de f_h par rapport au noyau $\mathcal{K}_h$. Ceci est un obstacle majeur à l'utilisation de la règle de Goldenshluger et Lepski (1.18) dans un cadre plus général de M-estimation. La majeure contribution du Chapitre 3 est de proposer une alternative à (1.18) permettant de traiter le cas des estimateurs non linéaires. Le principe est de comparer les risques empiriques (ou les gradients pour obtenir des vitesses de convergence rapides) à la place des estimateurs comme habituellement. Ainsi, il suffit que le risque empirique soit une fonctionnelle linéaire du noyau pour obtenir des résultats adaptatifs optimaux.

1.5 Description synthétique des différents chapitres

A présent, nous listons les contributions par chapitre. Notez que le Chapitre 2 est un aperçu des travaux [L3], [L8], [L4], [L15], [L10] et [L6], le Chapitre 3 est tiré de [L6], [L7] et [L12] alors que le Chapitre 4 traite des problèmes de suites individuelles ([L9] et [L11]). Le Chapitre 5 décrit des collaborations avec les biologistes, médecins et industriels ([L5],[L13] et [L17]).

Les résultats du Chapitre 2

Le chapitre 2 est un survol des principaux résultats obtenus dans le modèle d'apprentissage statistique de problèmes inverses. Dans une première section, nous cherchons à obtenir des vitesses de convergence minimax en analyse discriminante avec erreurs dans les variables. Dans un premier temps, deux bornes inférieures sont établies, généralisant les précédents travaux de Mammen and Tsybakov [1999] et Audibert and Tsybakov [2007] au modèle avec erreurs dans les variables. Les vitesses qui apparaissent dans ces bornes inférieures dépendent de la marge et de la complexité (comme précédemment dans le cas direct de Mammen and Tsybakov [1999] et Audibert and Tsybakov [2007]) et du degré du problème inverse, comme habituellement dans les modèles avec erreurs dans les variables. La suite de cette section s'intéresse à la construction d'estimateurs atteignant ces bornes. Pour cela, on propose de remplacer la densité des observations dans le risque par un estimateur à noyau de déconvolution, et de minimiser le risque empirique ainsi construit. Par la suite, nous proposons de généraliser ces bornes d'excès de risque dans plusieurs directions. Nous considerons un problème inverse d'apprentissage dans toute sa généralité, où l'on veut minimiser un risque $R_P(\cdot)$ à partir d'observations indirectes (Z_i, Y_i), $i = 1, \ldots, n$ de loi $(Z, Y) \sim \tilde{P}$, où $Z \sim Af$ avec A un opérateur linéaire compact et f la densité de X où $(X, Y) \sim P$. Dans ce cadre, nous obtenons des bornes d'excès de risque avec grande probabilité dans des modèles supervisés et non supervisés. Des contraintes anisotropes sur la densité f sont aussi étudiées dans un cadre non supervisé avec erreur dans les variables. Dans cette section, les techniques de localisation sont adaptées au cas indirect pour obtenir des vitesses de convergence rapides qui généralisent les résultats de Koltchinskii [2006] ou plus récemment Levrard [2013]. Dans une troisième section, en utilisant les résultats précédents, on construit un algorithme de type k-means pour le problème de clustering avec erreurs dans les variables. L'algorithme proposé imite l'algorithme de Lloyd (méthode de Newton), avec une étape de déconvolution avant le schéma itératif. Cet algorithme, appelé noisy k-means est testé sur des mélanges de gaussiennes avec erreur dans les variables. On illustre une bonne robustesse de notre méthode lorsque le bruit augmente, ce qui n'est pas le cas de l'algorithme des k-means.

Les résultats du Chapitre 3

Le Chapitre 3 est dédié au problème du choix de la fenêtre dans le problème de minimisation d'un risque empirique dépendant d'un paramètre. Dans un premier temps, le modèle de clustering avec erreurs dans les variables est étudié. Une vitesse adaptative rapide est obtenue pour une méthode de sélection de fenêtre isotrope $h \in \mathbb{R}$. Cette méthode, appelée ERC (Empirical Risk Comparison), est basée sur la méthode de Lepski (1.16) et remplace la comparaison d'estimateurs par une comparaison des risques empiriques. Les résultats obtenus sont des vitesses rapides adaptatives pour l'excès de risque. Enfin, pour être complet, nous proposons une nouvelle règle de sélection de fenêtre dans un cadre anisotrope. Pour cela, nous introduisons un nouveau critère qui mesure la norme du gradient du risque d'un estimateur. Comme présenté ci-dessus, ce critère permet d'obtenir des vitesses rapides en apprentissage statistique sans hypothèse de marge et sans technique de localisation, à condition que la fonction de perte soit suffisamment régulière. Ainsi, nous pouvons proposer une méthode de sélection basée sur la minimisation d'un majorant du compromis biais variance. Ce majorant est construit à l'aide d'un risque empirique auxiliaire, en suivant les idées de Goldenshluger and Lepski [2011]. Cette fois-ci, nous remplaçons la comparaison d'estimateurs par la comparaison des gradients. Les résultats sont obtenus dans un cadre général de minimisation d'un risque empirique dépendant d'une fenêtre, et permettent d'obtenir pour la première fois des vitesses adaptatives optimales pour des estimateurs non linéaires dans un cadre anisotrope.

Les résultats du chapitre 4

Le Chapitre 4 présente des résultats récents en apprentissage en ligne. Le jeu séquentiel introduit dans un premier temps est le suivant : à chaque tour $t = 1, \ldots, T$, à partir des observations passées $z_1, \ldots, z_{t-1} \in \mathbb{R}^d$, et sans avis d'experts, on veut prédire la position de z_t. Pour cela, on s'autorise plusieurs tentatives et la perte à l'instant t est la plus petite distance entre z_t et la tentative la plus proche. Ce problème de prédiction s'apparente à un problème de clustering en ligne. On démontre ainsi des bornes de regret sans aucun a priori sur le nombre de classes, ni aucun avis d'experts. Dans cette partie, la théorie PAC-Bayésienne en grande dimension permet d'obtenir un algorithme automatique qui sélectionne le nombre de groupes à chaque itération. Ces résultats sont convertis au cadre statistique traditionnel où des bornes d'excès de risque sont établies dans le cadre de la sélection de modèles (ici le nombre de classes) et du clustering en grande dimension. Par la suite, on étend ces résultats à un cadre de bi-clustering où le clustering est une étape intermédiaire dans un problème de prédiction en ligne. La dernière partie du Chapitre 4 étudie les propriétés d'optimalité des algorithmes séquentiels de clustering en ligne. Grâce à des outils probabilistes, on propose un comportement asymptotique du regret minimax dans le problème de clustering en ligne. Ces résultats illustrent l'optimalité d'une version pénalisée de l'algorithme de clustering de suites individuelles et de nombreux problèmes restent ouverts.

Les résultats du Chapitre 5

Ce chapitre expose plusieurs collaborations avec des scientifiques du vivant ou de l'industrie consistant à utiliser des techniques d'apprentissage pour résoudre des problèmes réels. Dans un premier temps, une synthèse des travaux en collaboration avec l'équipe GenHort (génétique et horticulture) de l'I.N.R.A. d'Angers et Koji Kawamura (biologiste à l'Osaka Institute of Technology) sur la recherche de QTL (Quantitative Trait Locus) est proposée. Dans ce problème, nous cherchons à déterminer des endroits du génome qui expliquent des variations phénotypiques de croisements de rosiers, et plus précisément des caractères d'architecture d'inflorescence. L'utilisation de méthodes à noyaux issues de l'apprentissage statistique (SVM, kernel PCA) s'avèrent très utile à la détection de ces QTL. Dans un second temps, j'aborde un projet de prédiction en médecine avec le C.H.U. d'Angers concernant le développement de méthodes non-invasives pour le diagnostic de la fibrose du foie. Dans ce problème, nous disposons d'un échantillon d'un millier de patients soumis à une analyse sanguine et à une biopsie du foie. L'objectif est d'agréger des régressions logistiques sur ces marqueurs sanguins pour prédire le stade de la fibrose. Enfin, la dernière partie est dédiée à un projet de prédiction dans le domaine du sport. L'entreprise Itnoveo développe une application Androïd pour partager l'évolution d'une rencontre sportive sur internet. En utilisant les SVM (Support Vector Machines), nous avons ajouté un plug-in de prédiction qui permet de pronostiquer le gagnant du match pendant la rencontre. Ces collaborations illustrent la diversité des applications potentielles des algorithmes d'apprentissage.

Notations

Chapter 2 and Chapter 3 could be group together to form a contribution to the statistical learning theory. We try, as far as possible, to use the same notations along these chapters according to the following table.

$X_i,\ i = 1, \ldots, n$	direct observations
$Z_i,\ i = 1, \ldots, n$	indirect observations
$\mathcal{K}(\cdot)$	kernel function
$\mathcal{F}[\cdot]$	Fourier transform
λ	generic smoothing parameter
$h \in \mathbb{R}_+^d$	bandwidth parameter
$\mathcal{H}$	bandwidth set
$\mathcal{K}_h(\cdot)$	h-dilation of a kernel function
$\widetilde{\mathcal{K}}_h(\cdot)$	deconvolution kernel
$N \in \mathbb{N}^*$	spectral cut-off
$R(\cdot)$	true risk
$\widehat{R}(\cdot)$	empirical risk
$\widehat{R}_h(\cdot)$	deconvoluted empirical risk
$R_h(\cdot)$	expectation of the deconvoluted empirical risk
$\mathbf{1}_G$	indicator function of the set G
$\mathcal{H}_B(\mathcal{G}, \epsilon, d)$	ϵ-entropy with bracketing of the set $\mathcal{G}$
$\mathbf{c}$	a codebook $(c_1, \ldots, c_k) \in \mathbb{R}^{dk}$
$\lvert \cdot \rvert_2,\ \lVert \cdot \rVert_2$	Euclidean norms

–Table 1. Notations for Chapter 2 and Chapter 3–

Chapitre 2

Inverse Statistical Learning

In this chapter, we undertake a survey of my contribution in statistical learning with indirect observations. The problem of indirect observations has been investigated for a while in nonparametric statistics. A patent example is density estimation in the presence of *noisy observations* :

$$Z_i = X_i + \epsilon_i,\ i = 1 \dots, n. \tag{2.1}$$

In this framework, $(X_i)_{i=1}^n$ are usually i.i.d. with unknown density f over $\mathbb{R}^d$ with respect to the Lebesgue measure whereas $(\epsilon_i)_{i=1}^n$ are i.i.d. with known density η, and independent of the sequence $(X_i)_{i=1}^n$. The purpose is to estimate f giving noisy observations $(Z_i)_{i=1}^n$. Simple Fourier analysis tell us that a good candidate estimator $\hat{f}$ should satisfy, under standard assumptions that will be examined in the sequel :

$$\mathcal{F}[\hat{f}] = \mathcal{F}[f_Z]/\mathcal{F}[\eta],$$

where f_Z is the density of Z_1 and $\mathcal{F}$ is the usual Fourier transform. In kernel estimation - which can be traced back to the work of Parzen (see Parzen [1962]) and Rosenblatt (see Rosenblatt [1956]) - this equation motivates the introduction of a deconvolution kernel[1] :

$$\widetilde{\mathcal{K}}_h(x) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}_h]}{\mathcal{F}[\eta]}\right](x), \tag{2.2}$$

where $\mathcal{K}_h(\cdot) = 1/h\mathcal{K}(\cdot/h)$ is the h-dilation of a given kernel $\mathcal{K}$. It leads to the following deconvolution kernel estimator :

$$\hat{f}_h(\cdot) = \frac{1}{n}\sum_{i=1}^n \widetilde{\mathcal{K}}_h(Z_i - \cdot). \tag{2.3}$$

The empirical mean (2.3) is focal in statistical learning with noisy observations (2.1). Roughly speaking, along the present dissertation, we recommend to plug (2.3) into the true risk of the problem.

On top of that, we will consider in the sequel a more general setting where noisy observations (2.1) are replaced by *indirect observations* Z_i, $i = 1, \dots, n$ with density Af where A is a known linear compact operator. In such a problem, (2.3) is replaced by another regularization scheme, such as projection or spectral cut-off (see Section 2.2) :

$$\hat{f}_N(\cdot) = \sum_{k=1}^N \hat{\theta}_k \phi_k(\cdot), \tag{2.4}$$

where $(\phi_k)_{k\in\mathbb{N}}$ is the orthonormal basis associated with the Singular Value Decomposition (SVD) of operator A whereas $(\hat{\theta}_k)_{k=1}^N$ are empirical coefficients.

This chapter is organized as follows. We bring up a precise minimax study of the problem of discriminant analysis with errors in variables in the first section. It combines four lower bounds, with related

[1] We adopt this notation from Chapter 2 to Chapter 3 where we need in particular the following property : $\widetilde{\mathcal{K}_h * \mathcal{K}_{h'}} = \widetilde{\mathcal{K}_h} * \widetilde{\mathcal{K}_{h'}}$, where $*$ stands for the convolution product.

upper bounds corresponding to different regularity assumptions, and at the same time two different criteria (see the definition of $d_{f,g}(\cdot,\cdot)$ and $d_\Delta(\cdot,\cdot)$ below). Loosely speaking, classification can be seen as a pure set estimation problem, whereas an alternative is to minimize an excess risk which takes into account the inherent difficulty of the problem. The regularity assumptions advanced in the sequel are :

— boundary fragment assumptions, initiated by Korostelëv and Tsybakov [1993] (see also Mammen and Tsybakov [1999]),

— and plug-in assumptions (see Yang [1999] or Audibert and Tsybakov [2007]).

Interestingly, these two regularity assumptions lead to quite different situations in the presence of indirect observations. The main message of this minimax study is the following one : when we observe a noisy sample as in (2.1), it is natural, easier, and in the end minimax to deal with plug-in type assumptions rather than boundary assumptions.

We continue this theory by extending the previous results to the general framework of statistical learning. In Section 2.2, we debate several risk bounds for smooth losses and common entropy conditions on the set of decision rules. We examine the problem of multiclass as well as unsupervised classification. We also extend the deconvolution framework to a general linear inverse problem as it was mentioned above. These generalizations illustrate rather well the convenience to consider a variety of problems with indirect observations. The introduction of universal complexity and margin assumptions enables us to extend the results of the noise free case (Koltchinskii [2006], see also Bartlett, Bousquet, and Mendelson [2005], Tsybakov [2004]).

These considerations permit to construct in Section 2.3 a novel algorithm for clustering with errors in variables. We focus on the prevailing k-means problem, where we want to learn k clusters of a set of observations. Following the guiding thread of the familiar Lloyd algorithm (Lloyd [1982]), we suggest a deconvoluted version of this algorithm called noisy k-means. Based on Newton's iteration, it is analyzed in several gaussian mixtures, and compared with a basic k-means algorithm.

2.1 Minimax theory [L3],[L8]

2.1.1 Introduction

In the problem of discriminant analysis, we usually observe two i.i.d. samples $X_1^{(1)},\dots,X_n^{(1)}$ and $X_1^{(2)},\dots,X_m^{(2)}$. Each observation $X_i^{(j)}\in\mathbb{R}^d$ is assumed to admit a density with respect to a σ-finite measure Q, dominated by the Lebesgue measure. This density will be denoted by f if the observation belongs to the first set (i.e. when $j=1$) or g in the other case. Our objective is to infer the density of a new incoming observation X. This problem can be seen as a particular case of the more general and extensively studied binary classification problem (see Devroye, Györfi, and Lugosi [1996] for a meticulous introduction or Boucheron, Bousquet, and Lugosi [2005] for a concise survey). In this framework, a decision rule or classifier can be identified with a set $G\subset\mathbb{R}^d$, which attributes X to f if $X\in G$ and to g otherwise. Then, we can associate to each classifier G its corresponding Bayes risk $R(G)$ defined as

$$R(G)=\frac{1}{2}\left[\int_{K\setminus G} f(x)dQ(x)+\int_G g(x)dQ(x)\right], \tag{2.5}$$

where we restrict the problem to a compact set $K\subset\mathbb{R}^d$. The minimizer of the Bayes risk (the best possible classifier for this criterion) is given by :

$$G_K^\star=\{x\in K : f(x)\geq g(x)\}, \tag{2.6}$$

where the infimum is taken over all subsets of K. The Bayes classifier is obviously unknown since it explicitly depends on the couple (f,g). The goal is thus to estimate $G_K^\star$ thanks to a classifier $\widehat{G}$ based on the two learning samples.

In the sequel, we propose two different measures of performances of a set $G\subset K$. First of all, simple algebra indicates that the excess risk $R(G)-R(G_K^\star)={}^1/_2\cdot d_{f,g}(G,G_K^\star)$ where $d_{f,g}(\cdot,\cdot)$ is a pseudo-distance over subsets of $K\subset\mathbb{R}^d$ defined as :

$$d_{f,g}(G_1,G_2)=\int_{G_1\Delta G_2}|f-g|dQ,$$

and $G_1 \Delta G_2 = [G_1^c \cap G_2] \cup [G_2^c \cap G_1]$ is the symmetric difference between two sets G_1 and G_2. In this context, there is another natural way of measuring the accuracy of a decision rule G through the quantity :

$$d_\Delta(G, G_K^\star) = \int_{G \Delta G_K^\star} dQ,$$

where d_Δ defines also a pseudo-distance on the subsets of $K \subset \mathbb{R}^d$.

In the noise free case, i.e. when $\epsilon = 0$ in (2.1), Mammen and Tsybakov [1999] has attracted the attention on minimax *fast rates of convergence* (i.e. faster than $n^{-1/2}$) and states in particular [2] :

$$\inf_{\hat{G}} \sup_{G_K^\star \in \mathcal{G}(\alpha,\rho)} \left[d_{f,g}(\hat{G}, G_K^\star) \right] \approx n^{-\frac{\alpha+1}{2+\alpha+\rho\alpha}}, \text{ as } n \to +\infty, \tag{2.7}$$

where $\mathcal{G}(\alpha, \rho)$ is a nonparametric set of candidates $G_K^\star$ with complexity $\rho > 0$ and margin parameter $\alpha \geq 0$ (see below for a precise definition). In (2.7), the complexity parameter $\rho > 0$ is associated to the notion of entropy with bracketing whereas the margin parameter is used to link the variance to the expectation. It gives Mammen and Tsybakov [1999] the opportunity to get improved bounds using the so-called peeling technique of van de Geer [2000]. This result is at the origin of a vast literature in classification (see for instance Massart and Nédélec [2006],Audibert and Tsybakov [2007], Blanchard, Bousquet, and Massart [2008], Blanchard, Lugosi, and Vayatis [2003]) or in general statistical learning (see Koltchinskii [2006], Bartlett, Bousquet, and Mendelson [2005], Bartlett and Mendelson [2006]). In these papers, the complexity assumption can be a geometric assumption over the class of candidates $G_K^\star$ (such as finite VC dimension, or boundary fragments) or hypotheses on the regularity of the regression function of classification (plug-in type assumptions). In Massart and Nédélec [2006], minimax fast rates are established for finite VC class of candidates whereas plug-in type assumptions have been studied in classification in Audibert and Tsybakov [2007] (see also Devroye, Györfi, and Lugosi [1996] or Yang [1999]). More generally, Koltchinskii [2006] proposes to consider $\rho > 0$ as a complexity parameter in local Rademacher complexities and gives universal upper bounds containing (2.7) and the results of Mammen and Tsybakov [1999] or Audibert and Tsybakov [2007].

In this section, we examine the estimation of the Bayes classifier $G_K^\star$ when dealing with noisy samples. For all $j \in \{1, 2\}$, we assume that we observe :

$$Z_i^{(j)} = X_i^{(j)} + \epsilon_i^{(j)}, \, i = 1, \dots n_j, \tag{2.8}$$

instead of the $X_i^{(j)}$, where in the sequel $n_1 = n$ and $n_2 = m$. The $\epsilon_i^{(j)}$ denotes random variables expressing measurement errors. We are facing an inverse problem, and more precisely a deconvolution problem. Deconvolution problems appear in many fields where data are obtained with measurements errors and are at the core of several nonparametric statistical studies. For a generous review of the possible methodologies associated to these problems, we may mention for instance Meister [2009]. More specifically, we refer to Fan [1991] in density estimation or Butucea [2007] where goodness-of-fit tests are brought up in the presence of noise. The key point of all these studies is to employ a deconvolution kernel in order to annihilate the noise ϵ. It is important to mention that in this discriminant analysis setup, or more conventionally in classification, there is - up to our knowledge - no such a work.

In the direct case, empirical risk minimizers appear as good candidates to reach fast rates of convergence. Unfortunately, in the error-in-variables model, since we observe noisy samples $Z = X + \epsilon$, classical ERM principle fails since :

$$\frac{1}{2n} \sum_{i=1}^{n} \mathbf{1}_{\{Z_i^{(1)} \in K \backslash G\}} + \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_{\{Z_i^{(2)} \in G\}} \longrightarrow \frac{1}{2} \left[\int_{K \backslash G} (f.\mu) * \eta(x) dx + \int_G (g.\mu) * \eta(x) dx \right] \neq R(G),$$

where $*$ stands for the convolution product (see below for details). This motivates a deconvolution step in the classical ERM procedure. We study the minimization of an asymptotically unbiased estimator

[2] where $u \approx v$ means that there exist $a, A > 0$ such that $av \leq u \leq Av$.

$\widehat{R}_h(G)$ of $R(G)$ which uses kernel deconvolution estimator (2.3) with bandwidth parameter h according to :

$$\widehat{R}_h(G) = \frac{1}{2n}\sum_{i=1}^{n} \widetilde{\mathcal{K}}_h * \mathbf{1}_{K\backslash G}(Z_i^{(1)}) + \frac{1}{2m}\sum_{i=1}^{m} \widetilde{\mathcal{K}}_h * \mathbf{1}_{G}(Z_i^{(2)}). \tag{2.9}$$

In this section, we set out as accurately as possible the influence of the error ϵ on the presence of fast rates of convergence. For this purpose, we apply the asymptotic theory of empirical processes in the spirit of van de Geer [2000] (see also van der Vaart and Wellner [1996]) to the deconvolution empirical risk (2.9). It leads to a new and interesting theory of risk bounds offered in Section 2.1.4 for discriminant analysis. In particular, we need to study in details the complexity of the class of functions $\{\widetilde{\mathcal{K}}_h * \mathbf{1}_G, G \in \mathcal{G}\}$. This complexity is related to the imposed complexity over $\mathcal{G}$, such as boundary fragment assumptions, or plug-in conditions. For each assumption, we establish lower and upper bounds and discuss the performances of this deconvolution ERM estimator for this problem. The problem of adaptation is postponed to Chapter 3 where adaptive fast rates of convergence are stated.

2.1.2 Plug-in (I) vs boundary fragments (II)

In this section, given a class $\mathcal{F}$, one would like to quantify as exactly as possible the minimax risks :

$$\inf_{\widehat{G}} \sup_{(f,g)\in\mathcal{F}} d_{\square}(\widehat{G}, G_K^\star),$$

where the infimum is taken over all possible estimators of $G_K^\star$ and $d_\square$ stands for $d_{f,g}$ or d_Δ. In order to obtain a satisfying minimax study, one needs to detail the considered classes $\mathcal{F}$. Such a class expresses some conditions that can be set on the pair (f, g). At the first glance, we detail some common assumptions (complexity and margin) that can be set on the pair (f, g). We then introduce the two main regularity assumptions, namely the plug-in and boundary fragments assumptions.

A first condition that can be set on the pair (f, g) is the well-known *margin assumption*. It has been introduced in discriminant analysis (see Mammen and Tsybakov [1999]) as follows :

Margin Assumption : *There exists positive constants $t_0, c_2, \alpha \geq 0$ such that for $0 < t < t_0$:*

$$Q\{x \in K : |f(x) - g(x)| \leq t\} \leq c_2 t^\alpha. \tag{2.10}$$

The margin assumption (2.10) is 'structural' in the sense that it describes the difficulty to distinguish an observation having density f from an other with density g. This assumption is related to the behaviour of $|f - g|$ at the boundary of $G_K^\star$. It may give a variety of minimax fast rates of convergence which depends on the margin parameter α. A large margin corresponds to configurations where the slope of $|f - g|$ is high at the boundary of $G_K^\star$. The most favorable case corresponds to a margin $\alpha = +\infty$ when $f - g$ jumps at the boundary of $G_K^\star$. The same kind of assumptions have been introduced originally in the related problem of excess mass by Polonik [1995].

From a statistical point of view, this assumption provides a precise description of the interaction between the pseudo distance $d_{f,g}$ and d_Δ. In particular, it permits a control of the variance of the empirical processes involved in the upper bounds. Other assumptions of this type can be formulated (see for instance Bartlett and Mendelson [2006] or Koltchinskii [2006]) in a more general statistical learning context. This is the focus of Section 2.2.2 (see for instance Definition 3).

For the sake of convenience, we will also require an additional hypothesis on the noise ϵ. We assume in the sequel that $\epsilon = (\epsilon_1, \ldots, \epsilon_d)'$ admits a density η with respect to the Lebesgue measure satisfying :

$$\eta(x) = \prod_{i=1}^{d} \eta_i(x_i)\ \forall x \in \mathbb{R}^d. \tag{2.11}$$

In other words, the entries of the vector ϵ are independent. The assumption below describes the difficulty of the problem. It is often called the ordinary smooth case in the inverse problem literature.

Noise Assumption : *There exist $(\beta_1, \ldots, \beta_d)' \in \mathbb{R}_+^d$ such that for all $i \in \{1, \ldots, d\}$, $\beta_i > 1/2$,*

$$|\mathcal{F}[\eta_i](t)| \sim |t|^{-\beta_i}, \text{ and } \left|\mathcal{F}'[\eta_i](t)\right| \sim |t|^{-\beta_i} \text{ as } t \to +\infty,$$

where $\mathcal{F}[\eta_i]$ denotes the Fourier transform of η_i. Moreover, we assume that $\mathcal{F}[\eta_i](t) \neq 0$ for all $t \in \mathbb{R}$ and $i \in \{1, \ldots, d\}$.

Classical results in deconvolution (see e.g. Fan [1991], Fan and Truong [1993] or Butucea [2007] among others) are stated for $d = 1$. Two different settings are then distinguished concerning the difficulty of the problem which is expressed through the shape of $\mathcal{F}[\eta]$. One considers alternatively the case where $|\mathcal{F}[\eta](t)| \sim |t|^{-\beta}$ as $t \to +\infty$, which yet corresponds to mildly ill-posed inverse problem or $|\mathcal{F}[\eta](t)| \sim e^{-\gamma t}$ as $t \to +\infty$ which leads to a severely ill-posed inverse problem. This last setting corresponds to a particularly difficult problem and is often associated to low minimax rates of convergence.

In this dissertation, we only deal with d-dimensional mildly ill-posed deconvolution problems. For the sake of brevity, we do not examine severely ill-posed inverse problems or possible intermediates (e.g. a combination of polynomial and exponential tails, see Comte and Lacour [2013]). Nevertheless, the rates in these cases could be obtained through the same steps.

In order to provide a complete study, one also needs to set an assumption on the difficulty to find $G_K^\star$ in a possible set of candidates, namely a complexity assumption. In the classification framework, two different kind of complexity assumptions are often proposed in the literature. The first one concerns the shape of $G_K^\star$, literally the regularity of the boundary of the Bayes classifier. Another way to describe the complexity of the problem is to impose condition on the regularity of the underlying densities f and g. Such kind of condition is originally related to plug-in approaches. Remark that any clear connexion can be established between these two assumptions : a set $G_K^\star$ with a smooth boundary is not necessarily associated to smooth densities.

We are now ready to give a precise description of these assumptions. In the following, we denote by $\Sigma(\gamma, L)$ the class of isotropic Hölder functions according to the following definition [3].

Definition 1. *Fix $\gamma > 0$ and $L > 0$, and let $\lfloor \gamma \rfloor$ be the largest integer strictly less than γ. The* isotropic Hölder class *$\Sigma(\gamma, L)$ on $K \subset \mathbb{R}^d$ is the set of functions $f : K \to \mathbb{R}$ having on K all partial derivatives of order $\lfloor \gamma \rfloor$ and such that for any $x, y \in K$:*

$$\left| \frac{\partial^{|p|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} - \frac{\partial^{|p|} f(y)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}} \right| \leq L \sum_{v=1}^{d} |x_v - y_v|^{\gamma - \lfloor \gamma \rfloor}, \quad \forall\, p \in \mathbb{N}^d \ : \ |p| := p_1 + \cdots + p_d = \lfloor \gamma \rfloor;$$

$$\sum_{m=0}^{\lfloor \gamma \rfloor} \sum_{|p|=m} \sup_{x \in \mathbb{R}^d} \left| \frac{\partial^{|p|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right| \leq L,$$

where x_v and y_v are the v^{th} components of x and y.

We are now on time to state the plug-in assumption as follows :

Plug-in Assumption (I). *There exists γ and L positive constants such that $f - g \in \Sigma(\gamma, L)$.*

This hypothesis concerns the regularity of the function $f - g$ itself. It allows to control the complexity of the set of candidates $G_K^\star$ and get minimax results for the problem of discriminant analysis with errors in variables (see Theorem 1 and Theorem 3 below).

Other conditions have been proposed in the literature in order to explain and quantify the difficulty related to a classification problem. We can consider a family of boundary fragments on $K = [0, 1]^d$ as

[3] The anisotropic case is not examined here for simplicity whereas it is the purpose of Section 2.2.3 (see Definition 5).

follows. A set $G \subset [0,1]^d$ belongs to a class of boundary fragments (see Korostelev and Tsybakov [1993]) if there exists $b : [0,1]^{d-1} \to [0,1]$ such that :

$$G = \{x = (x_1, \ldots x_d) : x_d \le b(x_1, \ldots, x_{d-1})\} := G_b.$$

For given $\gamma, L > 0$ the class of Hölder boundary fragments is then defined as

$$\mathcal{G}(\gamma, L) = \{G_b, b \in \Sigma'(\gamma, L)\}, \tag{2.12}$$

where $\Sigma'(\gamma, L)$ is here the class of $d-1$ isotropic Hölder functions on $K = [0,1]^{d-1}$.

Boundary fragment assumption (II). *There exist γ and L positive constants such that the set $G_K^\star$ belongs to $\mathcal{G}(\gamma, L)$.*

The boundary fragment assumption concerns the set $G_K^\star$ and in particular the smoothness of its boundary. This assumption allows to get minimax fast rates in the direct case (see Mammen and Tsybakov [1999]).

2.1.3 Lower bounds

Here we propose to state the two main lower bounds for the plug-in assumption (case I, see Theorem 1) and for the boundary fragment assumption (case II, see Theorem 2).

Lower bound I

We call $\mathcal{F}_{\text{plug}}(Q)$ the set of all pairs (f, g) satisfying both the *margin* (with respect to Q) and the *plug-in* assumptions, since the previous assumption is often associated to plug-in rules in the statistical learning literature. The following theorem proposes a lower bound in such a setting.

Theorem 1. *Suppose the Noise assumption is satisfied for some $\beta = (\beta_1, \ldots, \beta_d)^\top$. Then, there exists an absolutely continuous measure Q_0 such that provided $\alpha \le 1$,*

$$\liminf_{n \to +\infty} \inf_{\widehat{G}} \sup_{(f,g) \in \mathcal{F}_{\text{plug}}(Q_0)} (n \wedge m)^{\tau_d(\alpha,\beta,\gamma)} d_\Box(\widehat{G}, G_K^\star) > 0,$$

where the infinimum is taken over all possible estimators of the set $G_K^\star$ and

$$\tau_d(\alpha, \beta, \gamma) = \begin{cases} \dfrac{\gamma\alpha}{\gamma(2+\alpha) + d + 2\displaystyle\sum_{i=1}^d \beta_i} & \text{for } d_\Box = d_\Delta, \\[2em] \dfrac{\gamma(\alpha+1)}{\gamma(2+\alpha) + d + 2\displaystyle\sum_{i=1}^d \beta_i} & \text{for } d_\Box = d_{f,g}. \end{cases}$$

It is appealing to perceive that we obtain exactly the same lower bounds as Audibert and Tsybakov [2007] in the direct case, which yet corresponds to the situation where $\beta_j = 0$ for all $j \in \{1, \ldots, d\}$. In this particular framework, the minimax rate of convergence mainly depends on γ and α. As in other deconvolution problems, in the presence of errors in variables, the rates obtained in Theorem 1 are deteriorated. The price to pay is an additional term of the form $2\sum_{i=1}^d \beta_i$. This term clearly connects the difficulty of the problem to the tail behavior of the characteristic function of the noise distribution. This price to pay is comparable with existing results in density estimation, regression with errors in variables or goodness-of-fit testing. Last step is to get a corresponding upper bound to validate this lower bound in the presence of noise in variables.

Remark that this lower bound is valid only for $\alpha \le 1$. Since we use in the proof of Theorem 1 an algebra based on standard Fourier analytical tools, we have to consider sufficient smooth objects. As a consequence in the lower bounds, we can check the margin assumption only for values of $\alpha \le 1$. Nevertheless, we conjecture that this restriction is only due to technical reasons and that our result remains pertinent for all $\alpha, \gamma \in \mathbb{R}$ (see the open problems at the end of the manuscript).

Proof: The proof mixes standard lower bounds arguments from classification (see Audibert [2004] and Audibert and Tsybakov [2007]), but then uses some techniques which are specific to the inverse problem literature (see for instance Butucea [2007] or Meister [2009]). Beforehand, for all estimator $\hat{G}_{n,m}$ of the set $G_K^\star$, we have :

$$\sup_{(f,g)\in\mathcal{F}_{plug}(Q_0)} \mathbb{E}_{f,g} d_\Delta(\hat{G}_{n,m}, G_K^\star) \geq \sup_{f\in\mathcal{F}_1(Q_0)} \mathbb{E}_{g_0}\left[\mathbb{E}_f\left\{d_\Delta(\hat{G}_{n,m}, G_K^\star)|Z_1^{(2)},\dots,Z_m^{(2)}\right\}\right], \tag{2.13}$$

where in (2.13), the existence of the triplet $(Q_0, \mathcal{F}_1(Q_0), g_0)$ is operated thanks to the following lemma.

Lemma 1. *For any $\gamma > 0$, provided that $\alpha \leq 1$, there exists a triplet $(Q_0, \mathcal{F}_1(Q_0), g_0)$ such that :*

1. *$\mathcal{F}_1(Q_0) = \{f_{\overrightarrow{\sigma}}, \overrightarrow{\sigma} \in \{0,1\}^k\}$ is a finite class of densities with respect to a specific measure Q_0 and g_0 a fixed density with respect to Q_0 ;*
2. *$(f_{\overrightarrow{\sigma}}, g_0) \in \mathcal{F}_{\text{plug}}(Q_0)$ for all $\overrightarrow{\sigma} \in \{0,1\}^k$;*
3. *For $j = 0, 1$, if we denote by $\mathbb{P}_j \sim Z = X+\epsilon$ where $X \sim f_{\overrightarrow{\sigma}_j}$ and $\overrightarrow{\sigma}_j = (\sigma_1, \dots, \sigma_{j-1}, j, \sigma_{j+1}, \dots, \sigma_k)$ whereas $\epsilon \sim \eta$ satisfying the Noise assumption, we have :*

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) \leq C \times k^{-\zeta_d(\alpha,\gamma,\beta)} \text{ where } \zeta_d(\alpha,\gamma,\beta) = \gamma + \frac{\alpha\gamma + d}{2} + \sum_{i=1}^{d} \beta_i.$$

Now, thanks to an Assouad lemma for classification (see Audibert [2004]), we have :

$$\sup_{\overrightarrow{\sigma}\in\{0,1\}^k} \mathbb{E}_{f_{\overrightarrow{\sigma}}}\left\{d_\Delta(\hat{G}_{n,m}, G_K^*)|Z_1^{(2)},\dots,Z_m^{(2)}\right\} \geq \sum_{j=1}^{k}\left[1 - \sqrt{(1+\chi^2(\mathbb{P}_1,\mathbb{P}_0))^n - 1}\int_{B_j} dQ_0(x)\right] \geq c' k^{-\alpha\gamma/2},$$

provided that $B_j \subset K$ is a painstaking subset and $\chi^2(\mathbb{P}_1, \mathbb{P}_0) \leq {}^{C}/_{n}$ to have the last inequality. Then, the third assertion of Lemma 1 and a choice of $k = n^{1/\zeta_d(\alpha,\gamma,\beta)}$ allows to concludes the proof of the lower bound. ∎

Lower bound II

In the following, we denote by $\mathcal{F}_{\text{frag}}(Q)$ the set of all pairs (f, g) satisfying both the *margin* and *boundary fragment* assumptions. Theorem 2 states lower bounds for the minimax risks over the class $\mathcal{F}_{\text{frag}}(Q)$.

Theorem 2. *Suppose that Q is the Lebesgue measure on $K = [0,1]^d$ and that the Noise assumption is satisfied. Then :*

$$\liminf_{n\to+\infty} \inf_{\widehat{G}} \sup_{(f,g)\in\mathcal{F}_{\text{frag}}(Q)} (n \wedge m)^{\tau_d'(\alpha,\beta,\gamma)} d_\Box(\widehat{G}, G_K^\star) > 0,$$

where the infinimum is taken over all possible estimators of the set $G_K^\star$ and

$$\tau_d'(\alpha,\beta,\gamma) = \begin{cases} \dfrac{\gamma\alpha}{\gamma(2+\alpha) + (d-1)\alpha + 2\alpha\sum_{i=1}^{d-1}\beta_i + 2\alpha\beta_d\gamma} & \text{for } d_\Box = d_\Delta, \\[2ex] \dfrac{\gamma(\alpha+1)}{\gamma(2+\alpha) + (d-1)\alpha + 2\alpha\sum_{i=1}^{d-1}\beta_i + 2\alpha\beta_d\gamma} & \text{for } d_\Box = d_{f,g}. \end{cases}$$

Remark that we obtain exactly the same lower bounds as Mammen and Tsybakov [1999] when $\beta_i \equiv 0$. As in Theorem 1, the minimax rates of convergence mainly depend on γ and α. In the presence of noise in the variables, the rates obtained in Theorem 2 are slower. The price to pay is an additional term of the form $2\alpha\sum_{i=1}^{d-1}\beta_i + 2\alpha\beta_d\gamma$. This term clearly connects the difficulty of the problem to the values of the coefficients $\beta_1, \dots, \beta_d$. Moreover, the above expression highlights a connection between the margin parameter and the ill-posedness. The role of the margin parameter over the inverse problem can be summarized as follows. Higher is the margin, higher is the price to pay for a given degree of ill-posedness. When the margin parameter is small, the problem is difficult at the boundary of $G_K^\star$ and we can only expect a non-sharp estimation of $G_K^\star$. In this case, it is not significantly worst to add noise. On the contrary, for large margin parameter, there is nice hope to give a sharp estimation of $G_K^\star$ and then perturb the input variables have strong consequences in the performances. Eventually, in the above expression, the first $d-1$ components of ϵ have not the same impact as the last (vertical) component. This is due to the Hölder boundary fragment assumption.

Proof: The proof of Theorem 2 folllows the same lines as the proof of Theorem 1. Nonetheless, in this case, the boundary assumption makes the edifice of the lower bound easier. More exactly, a similar result as Lemma 1 could be offered without any restriction on the margin parameter $\alpha \geq 0$. ∎

2.1.4 Upper bounds

In the noise free case ($\epsilon_i^{(j)} = (0, \dots, 0)$ for all $i \in \{1, \dots, n\}, j \in \{1, 2\}$), we deal with two samples having respective densities f and g. We know for instance from Mammen and Tsybakov [1999] that in this case, ERM estimators reach the minimax rates of convergence when $\mathcal{G} = \mathcal{G}(\gamma, L)$ corresponds to the set of boundary fragments with $\gamma > d-1$. For larger set $\mathcal{G}(\gamma, L)$, the minimization can be restricted to a $\delta-$net of $\mathcal{G}(\gamma, L)$. With an additional assumption over the approximation power of this $\delta-$net, the same minimax rates can be achieved in a subset of $\mathcal{G}(\gamma, L)$. If we consider complexity assumptions related to the smoothness of $f-g$, we can show easily with Audibert and Tsybakov [2007] that an hybrid plug-in/ERM estimator attains the minimax rates of convergence in the noise free case. The principle of the method is to consider an empirical minimization over a particular class :

$$\mathcal{G} = \{\{f-g \geq 0\}, f-g \in \mathcal{N}_{n,m}\},$$

where $\mathcal{N}_{n,m}$ is a well-chosen $\delta-$net. With such a procedure, minimax rates can be obtained with no restriction over the parameter γ, α and d.

In noisy discriminant analysis, ERM estimator is no longer available. Hence, we need an additional deconvolution step. In this context, we can put forward a deconvolution kernel, provided that the noise has a non null Fourier transform, as expressed in the *Noise Assumption*. Such an assumption is rather obvious in the inverse problem literature (see e.g. Fan [1991], Butucea [2007] or Meister [2009]).

Let $\mathcal{K} = \prod_{j=1}^d \mathcal{K}_j : \mathbb{R}^d \to \mathbb{R}$ be a d-dimensional function defined as the product of d unidimensional functions $\mathcal{K}_j$. The properties of $\mathcal{K}$ leading to satisfying upper bounds will be precised later on. Then, if we denote by $h = (h_1, \dots, h_d)$ a set of (positive) bandwidths and by $\mathcal{K}_h(x) = \prod_{i=1}^d h_i^{-1} \mathcal{K}(x_1/h_1, \dots, x_d/h_d)$, we define the *deconvolution kernel* $\widetilde{\mathcal{K}}_h$ as :

$$\begin{aligned} \widetilde{\mathcal{K}}_h \ &: \ \mathbb{R}^d \to \mathbb{R} \\ & \quad t \mapsto \widetilde{\mathcal{K}}_h(t) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}_h](\cdot)}{\mathcal{F}[\eta]}\right](t), \end{aligned} \tag{2.14}$$

where $\mathcal{F}[\cdot]$ stands for the Fourier transform. Observe that $\widetilde{\mathcal{K}}_h$ depends on the distribution of ϵ through η which is supposed to be known. In this context, for all $G \subset K$, the risk $R(G)$ can be estimated by :

$$\widehat{R}_h(G) = \frac{1}{2}\left[\frac{1}{n}\sum_{j=1}^n \widetilde{\mathcal{K}}_h * \mathbf{1}_{K \setminus G}(Z_j^{(1)}) + \frac{1}{m}\sum_{j=1}^m \widetilde{\mathcal{K}}_h * \mathbf{1}_G(Z_j^{(2)})\right],$$

where for a given $G \subset K$ and $z \in \mathbb{R}^d$:

$$\widetilde{\mathcal{K}}_h * \mathbf{1}_G(z) = \int_G \widetilde{\mathcal{K}}_h(z-x)\,dx. \tag{2.15}$$

The empirical risk $\widehat{R}_h(\cdot)$, and its associated minimizer $\widehat{G}_h$ is at the core of the upper bounds. As a rule, following the pioneering's works of Vapnik (see Vapnik [2000]), we have for $R_h(\cdot) := \mathbb{E}\widehat{R}_h(\cdot)$:

$$\begin{aligned} R(\widehat{G}_h) - R(G_K^\star) &\leq R(\widehat{G}_h) - \widehat{R}_h(\widehat{G}_h) + \widehat{R}_h(G_K^\star) - R(G_K^\star) \\ &\leq R_h(\widehat{G}_h) - \widehat{R}_h(\widehat{G}_h) + \widehat{R}_h(G_K^\star) - R_h(G_K^\star) \\ &\quad +(R - R_h)(\widehat{G}_h) - (R - R_h)(G_K^\star) \\ &\leq \sup_{G\in\mathcal{G}} |R_h - \widehat{R}_h|(G, G_K^\star) + \sup_{G\in\mathcal{G}} |R_h - R|(G, G_K^\star), \end{aligned} \tag{2.16}$$

where we write for concision for any $G, G' \subset K$:

$$|R_h - \widehat{R}_h|(G, G') = |R_h(G) - R_h(G') - \widehat{R}_h(G) + \widehat{R}_h(G')|,$$

and similarly :

$$|R_h - R|(G, G') = |R_h(G) - R_h(G') - R(G) + R(G')|.$$

As a result, to get risk bounds, we have to deal with two opposing terms, namely a so-called variance term :

$$\sup_{G\in\mathcal{G}} |R_h - \widehat{R}_h|(G, G_K^\star), \tag{2.17}$$

and a bias term (since $\mathbb{E}\widehat{R}_h(G) \neq R(G)$) of the form :

$$\sup_{G\in\mathcal{G}} |R_h - R|(G, G_K^\star). \tag{2.18}$$

The variance term (2.17) gives rise to the study of increments of empirical processes. In Theorem 3-4 below, this control is based on entropy conditions and uniform concentration inequalities (see van de Geer [2000] or van der Vaart and Wellner [1996]). However, in the noisy case, empirical processes are indexed by a class of functions which depends on the smoothing parameter h. This is the major obstacle in the control of (2.17).

The bias term (2.18) is controlled by taking advantages of the properties of $\mathcal{G}$ and of the assumptions on the kernel $\mathcal{K}$. This bias term is inherent to the estimation procedure and can be (sometimes) connected to the standard bias term in nonparametric estimation. Lemma 2 below provides a simple way to control the bias term under the plug-in assumption.

Lemma 2. *Suppose $f - g \in \Sigma(\gamma, L)$. Suppose that the kernel $\mathcal{K}$ is of order $\lfloor\gamma\rfloor$. Then, we have, for some $C > 0$:*

$$\sup_{G\subset K} |R_h - R|(G) \leq C \sum_{i=1}^d h_i^\gamma.$$

PROOF: The proof is straightforward since we can write :

$$R(G) - \mathbb{E}R_h(G) = \int \mathbf{1}_{G^C}[(f-g) - \mathcal{K}_h * (f-g)]dQ.$$

Then, the control of the bias term is reduced to the control of the bias term in standard nonparametric density estimation, which gives in this isotropic case (see for instance Tsybakov [2004]), for some positive constant $c > 0$:

$$\sup_{x_0\in\mathbb{R}^d} |(f-g)(x_0) - \mathcal{K}_h * (f-g)(x_0)| \leq c \sum_{i=1}^d h_i^\gamma.$$

■

A slightly finer version of this lemma is used in the proof of Theorem 3, where we improve the RHS of Lemma 2 using the margin assumption.

The choice of h will be a trade-off between the two opposing terms (2.17) and (2.18). Small h leads to complex functions (2.15) and blasts the variance term whereas (2.18) vanishes when h tends to zero.

We are now ready to give asymptotic fast rates of convergence to validate the lower bound of Theorem 1. For this purpose, we will require the following assumption on the kernel $\mathcal{K}$ which appears in (2.14).

Kernel Assumption. *The Kernel $\mathcal{K}$ is such that $\mathcal{F}[\mathcal{K}]$ is bounded and compactly supported.*

The construction of smooth kernels satisfying the kernel assumption could be managed using for instance the so-called Meyer wavelet (see Mallat [2009]).

Upper bound I

For all $\delta > 0$, using the notion of entropy (see for instance van der Vaart and Wellner [1996]) for Hölderian function on compact sets, we can find a δ-network $\mathcal{N}_\delta$ on $\Sigma(\gamma, L)$ such that :

— $\log(\mathrm{card}(\mathcal{N}_\delta)) \leq A\delta^{-d/\gamma}$,

— For all $h_0 \in \Sigma(\gamma, L)$, we can find $h \in \mathcal{N}_\delta$ such that $\|h - h_0\|_\infty \leq \delta$.

In the following, we associate to each $\nu := f - g \in \Sigma(\gamma, L)$, a set $G_\nu = \{x \in K : \nu(x) \geq 0\}$ and define the ERM estimator as :

$$\widehat{G}_h = \arg\min_{\nu \in \mathcal{N}_\delta} \widehat{R}_h(G_\nu), \tag{2.19}$$

where $\delta = \delta_{n,m}$ has to be chosen carefully. This procedure has been introduced in the direct case by Audibert and Tsybakov [2007] and referred to an hybrid Plug-in/ERM procedure [4]. The following theorem describes the performances of $\widehat{G}_h$.

Theorem 3. *Let $\widehat{G}_h$ the set introduced in (2.19) with :*

$$h_j \equiv (n \wedge m)^{-\frac{1}{\gamma(2+\alpha)+2\sum_{i=1}^d \beta_i + d}},\ \forall j \in \{1, \dots, d\},\ \text{and}\ \delta = \left(\frac{\prod_{i=1}^d h_i^{-\beta_i}}{\sqrt{n \wedge m}}\right)^{\frac{2}{d/\gamma+2+\alpha}}.$$

Given some $\sigma-$finite measure Q, suppose $(f, g) \in \mathcal{F}_{\mathrm{plug}}(Q)$ and the Noise assumption is satisfied with $\beta_i > 1/2$, $\forall i = 1, \dots d$. Consider a deconvolution kernel $\widetilde{\mathcal{K}}_h$ defined as in (2.14) where $\mathcal{K} = \Pi_{j=1}^d \mathcal{K}_j$ is a kernel of order $\lfloor \gamma \rfloor$, which satisfies the Kernel assumption. Then, for all real $\alpha \geq 0$:

$$\lim_{n,m \to +\infty} \sup_{(f,g) \in \mathcal{F}_{\mathrm{plug}}(Q)} (n \wedge m)^{\tau_d(\alpha,\beta,\gamma)} \mathbb{E}_{f,g} d_\Box(\widehat{G}, G_K^\star) < +\infty,$$

where Q is the Lebesgue [5] mesure and :

$$\tau_d(\alpha, \beta, \gamma) = \begin{cases} \dfrac{\gamma\alpha}{\gamma(2+\alpha) + d + 2\displaystyle\sum_{i=1}^d \beta_i} & \text{for } d_\Box = d_\Delta \\[2ex] \dfrac{\gamma(\alpha+1)}{\gamma(2+\alpha) + d + 2\displaystyle\sum_{i=1}^d \beta_i} & \text{for } d_\Box = d_{f,g}. \end{cases}$$

[4] We do not study plug-in rules in this chapter. Such algorithms are characterized by classifiers of the form :

$$\tilde{G}_{n,m} = \left\{x \in K,\ \tilde{f}_n(x) - \tilde{g}_m(x) \geq 0\right\},$$

where $\tilde{f}_n - \tilde{g}_m$ is an (optimal) estimator of the function $f - g$. The performances of such kind of methods have been investigated by Audibert and Tsybakov [2007] in the binary classification model. We also mention for instance Goldstein and Messer [1992] or Bickel and Ritov [2003] for contributions in a more general framework.

[5] A slightly finer result is proposed in [L3] where a more general measure is proposed. We omit these considerations here for concision.

Theorem 3 validates the lower bounds of Theorem 1. Deconvolution ERM are minimax optimal over the class $\mathcal{F}_{\text{plug}}(Q)$.

Here, fast rates (i.e. faster than $1/\sqrt{n}$) are pointed out when $\alpha\gamma > d+2\sum\beta_i$. This result is comparable to Audibert and Tsybakov [2007], where fast rates are proposed when $\alpha\gamma > d$. However, it is important to stress that large values of both α and γ correspond to restrictive situations. In this case, the margin parameter is high whereas the behavior of $f-g$ is smooth, which seems to be contradictory (see the related discussion in Audibert and Tsybakov [2007]).

The choice of h in Theorem 3 is the trade-off between the variance term (2.17) and the bias term (2.18). It is interesting to remark that this choice for h does not correspond with the optimal choice in the problem of deconvolution estimation of $f-g\in\Sigma(\gamma,L)$ thanks to noisy data. Here, the bandwidth depends on the margin parameter α and optimizes the classification excess risk bound. It highlights that the estimation procedure (2.19) is not a plug-in rule but an hybrid ERM/Plug-in estimator as in Audibert and Tsybakov [2007].

The minimax optimality of the procedure (2.19) is based on the choice of h in Theorem 3. This choice depends on unknown parameters such as the regularity of the function $f-g$. In this direction, adaptive fast rates are proposed in Chapter 3 in a slightly different framework [6].

Eventually, a similar approach can be considered in the direct case, using standard kernel estimators instead of deconvolution kernel estimators. The following corollary provides a new minimax procedure in the direct case.

Corollary 1. *Let $\bar{G}_h := \arg\min_{\nu\in\mathcal{N}_\delta} \bar{R}_h(G_\nu)$ where $\bar{R}_h(\cdot)$ is defined as :*

$$\bar{R}_h(G) = \frac{1}{2}\left[\frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h * \mathbf{1}_{K\backslash G}(X_j^{(1)}) + \frac{1}{m}\sum_{j=1}^{m}\mathcal{K}_h * \mathbf{1}_{G}(X_j^{(2)})\right].$$

Then, if $\mathcal{K}=\Pi_{j=1}^{d}\mathcal{K}_j$ is a kernel of order $\lfloor\gamma\rfloor$ satisfying the kernel assumption, if we choose :

$$h_j \le (n\wedge m)^{-\frac{1}{\gamma(2+\alpha)+d}},\ \forall j\in\{1,\dots,d\},\ \text{and}\ \delta=\delta_{n,m}=\left(\frac{1}{\sqrt{n\wedge m}}\right)^{\frac{2}{d/\gamma+2+\alpha}},$$

for any real $\alpha\ge 0$:

$$\lim_{n,m\to+\infty}\ \sup_{(f,g)\in\mathcal{F}_{\text{plug}}(Q)} (n\wedge m)^{\tau_d(\alpha,\gamma)}\mathbb{E}d_\square(\bar{G}_h, G_K^\star) < +\infty,$$

where Q is the Lebesgue measure and :

$$\tau_d(\alpha,\gamma) = \begin{cases} \dfrac{\gamma\alpha}{\gamma(2+\alpha)+d} & \textit{for } d_\square = d_\Delta \\[2ex] \dfrac{\gamma(\alpha+1)}{\gamma(2+\alpha)+d} & \textit{for } d_\square = d_{f,g}. \end{cases}$$

The choice of h in Corollary 1 is not standard. If h is small enough, the ERM procedure is minimax since in this case, the kernel function behaves like the Dirac function. This idea has been already mentioned in Vapnik [2000] in the general learning context and called Vicinal Risk Minimization (see also Chapelle, Weston, Bottou, and Vapnik [2001]). However, up to our knowledge, minimax asymptotic rates of convergence for this empirical minimization principle have not been proposed.

Upper bounds II

Here, we try to generalize the result of Mammen and Tsybakov [1999] to the noisy setting. We are looking at upper bounds matching with the lower bounds of Theorem 2. For the sake of concision, in this

[6] An adaptive version of Theorem 3 can be considered in the isotropic case using empirical risk comparisons as in Chapter 3.

paragraph we propose to restrict the set $\mathcal{G}$ to $\mathcal{G}(\gamma, L)$, where all possible regularities γ satisfy $\gamma > d-1$. Therefore, we are interested in the performances of the estimator :

$$\widetilde{G}_h = \arg \min_{G \in \mathcal{G}(\gamma, L)} \widehat{R}_h(G), \tag{2.20}$$

where the infimum is taken over the whole set $\mathcal{G}(\gamma, L)$ for $\gamma > d-1$ (see (2.12)). We will also assume for clarity that $n = m$ in the following theorem.

Theorem 4. *Let $\widetilde{G}_h$ the set introduced in* (2.20) *where $\gamma > d-1$. Suppose that the Noise assumption is satisfied and consider a kernel $\widetilde{\mathcal{K}}_h(\cdot)$ satisfying :*

$$\sup_{t \in \mathbb{R}^d} \left| \mathcal{F}[\widetilde{\mathcal{K}}_h](t) \right| \leq C \prod_{i=1}^{d} \lambda_i^{-\beta_i - 1}, \text{and } \|\widetilde{\mathcal{K}}_h\|^2 \leq C \prod_{i=1}^{d} \lambda_i^{-2\beta_i - 1}.$$

Suppose moreover that for any $j \in \{1, \ldots, d\}$, $\int_{\mathbb{R}^d} |\mathcal{K}(z)||z_j|\, dz < \infty$ and $\Pi_{j=1}^{d-1} \mathcal{K}_j$ has compact support. Then, if Q is the Lebesgue measure :

$$\lim_{n \to +\infty} \sup_{(f,g) \in \mathcal{F}_{\text{frag}}(Q)} n^{\kappa_d(\alpha, \beta, \gamma)} d_{\square}(\widehat{G}, G_K^{\star}) < \infty$$

where

$$\kappa_d(\alpha, \beta, \gamma) = \begin{cases} \dfrac{\gamma \alpha}{\gamma(\alpha+2) + (d-1)\alpha + 2\gamma(\alpha+1) \displaystyle\sum_{i=1}^{d} \beta_i} & \textit{for } d_{\square} = d_{\Delta} \\ \dfrac{\gamma(\alpha+1)}{\gamma(\alpha+2) + (d-1)\alpha + 2\gamma(\alpha+1) \displaystyle\sum_{i=1}^{d} \beta_i} & \textit{for } d_{\square} = d_{f,g}. \end{cases}$$

Following Theorem 4, lower and upper bounds do not match. The prize to pay for the errors-in-variables model is summarized in the term $2\gamma(\alpha+1)\sum_{i=1}^{d} \beta_i$ whereas the lower bound proposes a smaller term $2\alpha \sum_{i=1}^{d-1} \beta_i + 2\gamma\alpha\beta_d$. By the way, the corresponding error becomes negligible when γ is close to 1 and $\alpha \to \infty$. At the end of the manuscript, we discuss several tracks to attack this problem.

Proof: The proof uses empirical process theory gathering with the following crude bound for the bias term :

$$\sup_{G \in \mathcal{G}} (R_h - R)(G) \leq C \sum_{i=1}^{d} h_i.$$

It is based on the following scheme. For all $G \subset K$, using Fubini, we have

$$\begin{aligned} &\int_{\mathbb{R}^d} (f-g)(x) \left(\mathcal{K}_h * \mathbf{1}_G(x) - \mathbf{1}_G(x) \right) dx \\ &= \int_{\mathbb{R}^d} (f-g)(x) \left(\int_{z \in \mathbb{R}^2} \mathcal{K}(z) \left[\mathbf{1}_G(x+hz) - \mathbf{1}_G(x) \right] dz \right) dx \\ &= \int_{\mathbb{R}^d} \mathcal{K}(z) \left(\int_{\mathbb{R}^d} (f-g)(x) \left[\mathbf{1}_G(x+hz) - \mathbf{1}_G(x) \right] dx \right) dz. \end{aligned}$$

Since we do not have any conditions on the smoothness of $f-g$, the control of the bias reduces to the calculation of the Lebesgue measure between the sets G and $G + hz$, which appears to be of order $\sum_i h_i$. Hence, we can not take advantage on the smoothness of the boundary. ■

2.2 Other risk bounds [L4],[L6],[L16]

In this section, we survey some generalizations of the setting of Section 2.1 in two directions :

— From the statistical learning point of view, we consider in the sequel a general loss function ℓ and a general set of decision rule $\mathcal{G}$. We forget the presence of two samples with two different densities and consider a couple $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ of random variables, where the label $Y \in \mathcal{Y}$. It allows to consider different well-known learning problems, from multiclass classification to clustering. My weakness for classification in general makes this section dedicated to these fields. However, we can emphasize that many other problems could be considered, such as anomaly detection, learning principal curves, level-set estimation or ranking, just to name a few. Furthermore, we propose more general complexity assumptions for the generic decision set $\mathcal{G}$, in terms of ϵ-entropy. The main message at this point is the following : a precise study of the empirical process we have at hand allows us to give right order upper bounds under standard entropy conditions. For this purpose, we use several localization techniques introduced in the last two decades by many authors. It turns out that local complexity, such as modulus of continuity of empirical processes, have to be controlled.

— We can also consider a generic linear inverse problem instead of a particular deconvolution problem related with the Fourier domain. To be honest, we do not have meet such a problem in practice, where errors-in-variables models are more commonly used. However, from a theoretical point of view, it is interesting to ask the following question : can we consider a general linear operator and other regularization schemes instead of kernel deconvolution estimators ?

The problem of inverse statistical learning

Let us consider a generator of random inputs $X \in \mathcal{X}$, with unknown density distribution f with respect to some σ-finite measure Q, and (a possible) associated output $Y \in \mathcal{Y}$, from an unknown conditional probability. The joint law of (X, Y) is denoted by P. Given a class of functions $g \in \mathcal{G}$, we suppose the existence of an oracle defined as :

$$g^\star \in \arg\min_{g \in \mathcal{G}} R(g), \tag{2.21}$$

where $R(g) := \mathbb{E}_P \ell(g, (X, Y))$ is the risk associated to a general loss function. For example, the set $\mathcal{G}$ can be functions $g : x \in \mathcal{X} \mapsto g(x) \in \mathcal{Y}$, with $\ell(g, (x, y)) = \Phi(y - g(x))$ a prediction loss function. The problem of *inverse statistical learning* consists in estimating the oracle $g^\star$ based on a set of *indirect observations* :

$$(Z_i, Y_i),\ i = 1 \ldots, n,\ \text{where } Z_i \sim Af, \tag{2.22}$$

with A a given linear compact operator. The density of the direct and unobserved input variable X is denoted by f, whereas the joint law of (Z, Y) is written $\tilde{P}$. We are facing an inverse problem. In the sequel, we consider a bounded loss function ℓ such that for any $g \in \mathcal{G}$, $\ell(g, \cdot) : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ and a compact input space $\mathcal{X} \subset \mathbb{R}^d$. Given a class $\mathcal{G}$ of measurable functions $g : \mathcal{X} \to \mathbb{R}$, the performances of a given g is measured through its non-negative excess risk, given by :

$$R(g) - R(g^\star),$$

where $g^\star$ is defined in (2.21) as a minimizer of the risk.

The most extensively studied model with indirect observations is the additive measurement error model (see Section 2.1). In this case, we observe indirect inputs $Z_i = X_i + \epsilon_i, i = 1, \ldots, n$, where $(\epsilon_i)_{i=1}^n$ are i.i.d. with known density η. It corresponds to a convolution operator $A_\eta : f \mapsto f * \eta$ in (2.22). Depending on the nature of the response $Y \in \mathcal{Y}$, we deal with classification with errors in variables, density deconvolution, or regression with errors in variables.

Here, given a linear compact operator A, we observe a corrupted sample $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ where $Z_i, i = 1, \ldots, n$ are i.i.d. with density Af. Following Section 2.1, in this general context, given a smoothing parameter λ, we consider a λ-Empirical Risk Minimization (λ-ERM for short in the sequel) :

$$\arg\min_{g \in \mathcal{G}} \widehat{R}_\lambda(g), \tag{2.23}$$

where $\widehat{R}_\lambda(g)$ is defined in a general way as :

$$\widehat{R}_\lambda(g) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(g,(x,y))\hat{P}_\lambda(dx,dy). \tag{2.24}$$

The random measure $\hat{P}_\lambda = \hat{P}_\lambda(Z_1,Y_1,\ldots,Z_n,Y_n)$ is data-dependent and uses standard regularization methods coming from the inverse problem literature (see Engl, Hank, and Neubauer [1996]). Explicit constructions of $\hat{P}_\lambda$ and empirical risk (2.24) are elaborated below. This construction depends on the inverse problem that we have at hand, and the regularization method used. Consequently, the smoothing parameter may be the bandwidth of some deconvolution kernel estimator (Section 2.1), or some threshold of a spectral cut-off. We denote it as λ in full generality.

2.2.1 A general upper bound in multiclass classification

In this paragraph, we put forward the construction of the empirical risk (2.24) in supervised classification, i.e. when $\mathcal{Y}=\{0,\ldots,M\}$ in (2.22) for some $M\geq 1$. We introduce minimal assumptions to control the expected excess risk of the procedure. The formation of the empirical risk is based on the following decomposition of the true risk :

$$R(g) = \sum_{y\in\mathcal{Y}} p(y)\int_{\mathcal{X}} \ell(g,(x,y))f_y(x)Q(dx), \tag{2.25}$$

where $f_y(\cdot)$ is the conditional density of X given $Y=y$ and $p(y)=\mathbb{P}(Y=y)$, for any $y\in\mathcal{Y}$. With such a decomposition, we suggest to replace each $f_y(\cdot)$ by a nonparametric density estimator. To state a general upper bound, given $n_y=\text{card}\{i=1,\ldots,n : Y_i=y\}$, $k_\lambda:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ and the set of inputs $(Z_i^y)_{i=1}^{n_y}=\{Z_i, i=1,\ldots,n : Y_i=y\}$, we consider a family of estimators satisfying :

$$\forall y\in\mathcal{Y}, \hat{f}_y(\cdot) = \frac{1}{n_y}\sum_{i=1}^{n_y} k_\lambda(Z_i^y,\cdot). \tag{2.26}$$

Equation (2.26) provides a variety of nonparametric estimators of $f_y(\cdot)$. For instance, in the minimax study of classification with errors-in-variables, we have constructed deconvolution kernel estimators. In this case, $k_\lambda(x,y)=\widetilde{\mathcal{K}}_h(y-x)$ and the smoothing parameter corresponds to the d-dimensional bandwidth of the deconvolution kernel. Another standard representation such as (2.26) is proposed in this section with projection estimators (or spectral cut-off) using the SVD of operator A. In this case, the smoothing parameter is the dimension of the projection method. Of course, many other regularization methods could be considered, such as Tikhonov regularization. Moreover, it is important to mention that (2.26) considers a constant smoothing level λ for any class $y\in\mathcal{Y}$. This could be relaxed.

As in Section 2.1 above, we plug (2.26) in the true risk to get an empirical risk defined as :

$$\widehat{R}_\lambda(g) = \sum_{y\in\mathcal{Y}}\int_{\mathcal{X}} \ell(g,(x,y))\hat{f}_y(x)Q(dx)\hat{p}(y),$$

where $\hat{p}(y)=\frac{n_y}{n}$ is an estimator of the quantity $p(y)=\mathbb{P}(Y=y)$. Thanks to (2.26), this empirical risk can be written as :

$$\widehat{R}_\lambda(g) = \frac{1}{n}\sum_{i=1}^{n} \ell_\lambda(g,(Z_i,Y_i)), \tag{2.27}$$

where $\ell_\lambda(g,(z,y))$ is a modified version of $\ell(g,(x,y))$ given by :

$$\ell_\lambda(g,(z,y)) = \int_{\mathcal{X}} \ell(g,(x,y))k_\lambda(z,x)Q(dx).$$

In order to state upper bounds for a minimizer of (2.27), we need the following definition [7].

[7] In the sequel, we write $\ell_\lambda(g):(x,y)\mapsto \ell_\lambda(g,(x,y))$.

Definition 2. *We say that the class $\{\ell_\lambda(g), g \in \mathcal{G}\}$ is a LB-class (Lipschitz bounded class) with respect to μ with parameters $(c(\lambda), K(\lambda))$ if these two properties hold :*

$(\mathbf{L}_\mu)$ *$\{\ell_\lambda(g), g \in \mathcal{G}\}$ is Lipschitz w.r.t. μ with constant $c(\lambda)$:*

$$\forall g, g' \in \mathcal{G},\ \|\ell_\lambda(g) - \ell_\lambda(g')\|_{L_2(\widetilde{P})} \leq c(\lambda)\|\ell(g) - \ell(g')\|_{L_2(\mu)}.$$

(B) *$\{\ell_\lambda(g), g \in \mathcal{G}\}$ is uniformly bounded with constant $K(\lambda)$:*

$$\sup_{g\in\mathcal{G}} \sup_{(z,y)} |\ell_\lambda(g,(z,y))| \leq K(\lambda).$$

A LB-class of loss function is Lipschitz and bounded with constants which depend on λ. Examples of LB-classes are presented in the sequel. Coarsely speaking, the dependence on λ is driven by the behaviour of the noise density η, as in the *Noise assumption* in Section 2.1. These properties are necessary to derive explicitly an upper bound of the variance as a function of λ. Eventually, we will see that the Lipschitz constant $c(\lambda)$ in Definition 2 summarizes exactly the price to pay for the inverse problem in the excess risk bounds.

In this way, the Lipschitz property $(\mathbf{L}_\mu)$ is a key ingredient to control the complexity of the class of functions $\{\ell_\lambda(g) - \ell_\lambda(g^\star), g \in \mathcal{G}\}$. In the sequel, we use the following geometric complexity parameter :

$$\widetilde{\omega}_n(\mathcal{G}, \delta, \mu) = \mathbb{E} \sup_{g,g'\in\mathcal{G}:\|\ell(g)-\ell(g')\|_{L_2(\mu)}\leq\delta} \left|(\widehat{R}_\lambda - R_\lambda)(g - g')\right|. \tag{2.28}$$

This quantity corresponds to the indirect counterpart of more classical local complexities introduced in a variety of papers (see Bartlett, Bousquet, and Mendelson [2005], Koltchinskii [2006], Massart [2000]). Its control as a function of n, δ and λ is central to get fast rates of convergence. This can be done thanks to the following lemma.

Lemma 3. *Consider a LB-class $\{\ell_\lambda(g), g \in \mathcal{G}\}$ with respect to μ with Lipschitz constant $c(\lambda)$. Then, given some $0 < \rho < 1$, we have for some $C_1 > 0$:*

$$\mathcal{H}_B(\{\ell(g),\, g \in \mathcal{G}\}, \epsilon, L_2(\mu)) \leq c\epsilon^{-2\rho} \Rightarrow \widetilde{\omega}_n(\mathcal{G}, \delta, \mu) \leq C_1 \frac{c(\lambda)}{\sqrt{n}} \delta^{1-\rho},$$

where $\mathcal{H}_B(\{\ell(g),\, g \in \mathcal{G}\}, \epsilon, L_2(\mu))$ denotes the ϵ-entropy with bracketing of the set $\{\ell(g),\, g \in \mathcal{G}\}$ with respect to $L_2(\mu)$ (see van der Vaart and Wellner [1996] for a definition).

With such a lemma, it is possible to control the complexity in the indirect setup thanks to standard entropy conditions related with the class $\mathcal{G}$. The proof is based on a maximal inequality due to van der Vaart and Wellner [1996] applied to the class :

$$\mathcal{F}_\lambda = \{\ell_\lambda(g) - \ell_\lambda(g') : \|\ell(g) - \ell(g')\|_{L_2(\mu)} \leq \delta\}.$$

Eventually, in Definition 2, **(B)** is also necessary to apply Bousquet's inequality. This condition could be relaxed by dint of recent advances on empirical processes in an unbounded framework (see Lecué and Mendelson [2012] or Lederer and van de Geer [2012]).

Another standard assumption to get fast rates of convergence is the so-called Bernstein assumption. It can be linked with the standard margin assumption introduced in discriminant analysis by Mammen and Tsybakov [1999] (see also Section 2.1).

Definition 3. *For $\kappa \geq 1$, we say that $\mathcal{F}$ is a Bernstein class with respect to μ with parameter κ if there exists $\kappa_0 \geq 0$ such that for every $f \in \mathcal{F}$:*

$$\|f\|^2_{L_2(\mu)} \leq \kappa_0 [\mathbb{E}_P f]^{\frac{1}{\kappa}}.$$

This assumption first appears in Bartlett and Mendelson [2006] for $\mu = P$ when $\mathcal{F} = \{\ell(g) - \ell(g^\star), g \in \mathcal{G}\}$ is the excess loss class. It allows to control the excess risk in statistical learning using functional's Bernstein inequality such as Talagrand's type inequality. In classification, it corresponds to the standard margin assumption (see Section 2), where in this case $\kappa = \frac{\alpha+1}{\alpha}$ for a so-called margin parameter $\alpha \geq 0$.

Definition 3 has to be combined with the Lipschitz property of Definition 2. It allows us to have the following serie of inequalities :

$$\|\ell_\lambda(g) - \ell_\lambda(g^\star)\|_{L_2(\widetilde{P})} \leq c(\lambda)\|f\|_{L_2(\mu)} \leq c(\lambda)\left(\mathbb{E}_P f\right)^{\frac{1}{2\kappa}}, \tag{2.29}$$

where $f \in \mathcal{F} = \{\ell(g) - \ell(g^\star),\, g \in \mathcal{G}\}$. Last definition provides a control of the bias term as follows :

Definition 4. *The class $\{\ell_\lambda(g), g \in \mathcal{G}\}$ has approximation function $a(\lambda)$ and residual constant $0 < r < 1$ if the following holds :*

$$\forall g \in \mathcal{G},\, (R - R_\lambda)(g - g^\star) \leq a(\lambda) + r(R(g) - R(g^\star)),$$

where with a slight abuse of notations, we write :

$$(R - R_\lambda)(g - g^\star) = R(g) - R(g^\star) - R_\lambda(g) + R_\lambda(g^\star).$$

This definition warrants a control of the bias term. It is straightforward that with Definition 4, gathering with a bias variance decomposition as in (3.6), we get a control of the excess risk of $\widehat{g}_\lambda$ defined in (2.30) as follows :

$$R(\widehat{g}_\lambda) - R(g^\star) \leq \quad R(\widehat{g}_\lambda) - \widehat{R}_\lambda(\widehat{g}_\lambda) + \widehat{R}_\lambda(g^\star) - R(g^\star) \leq \frac{1}{1-r}\left(\sup_{g \in \mathcal{G}(1)} |(\widehat{R}_\lambda - R_\lambda)(g - g^\star)| + a(\lambda)\right),$$

where in the sequel :

$$\mathcal{G}(\delta) = \{g \in \mathcal{G} : R(g) - R(g^\star) \leq \delta\}.$$

Explicit functions $a(\lambda)$ and residual constant $r < 1$ are obtained in applications. There depend on the regularity conditions over the conditional densities as well as the margin parameter $\kappa \geq 1$ in Definition 3. It allows to get fast rates of convergence. We are now on time to state the following general upper bound for the expected excess risk of the estimator :

$$\widehat{g}_\lambda \in \arg\min_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \ell_\lambda(g, (Z_i, Y_i)). \tag{2.30}$$

Theorem 5. *Suppose $\{\ell(g) - \ell(g^\star), g \in \mathcal{G}\}$ is Bernstein with respect to μ with parameter $\kappa \geq 1$ where $g^\star = \arg\min_{\mathcal{G}} R(g)$ is unique*[8]*. Suppose there exists $0 < \rho < 1$ such that :*

$$\mathcal{H}_B(\{\ell(g),\, g \in \mathcal{G}\}, \epsilon, L_2(\mu)) \leq C_2 \epsilon^{-2\rho}, \tag{2.31}$$

for some $C_2 > 0$.
Consider a LB-class $\{\ell_\lambda(g), g \in \mathcal{G}\}$ with respect to μ with parameters $(c(\lambda), K(\lambda))$ and approximation function $a(\lambda)$ such that :

$$a(\lambda) \leq C_1 \left(\frac{c(\lambda)}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}} \text{ and } K(\lambda) \leq \frac{c(\lambda)^{\frac{2\kappa}{2\kappa+\rho-1}} n^{\frac{\kappa+\rho-1}{2\kappa+\rho-1}}}{1 + \log\log_q n}, \tag{2.32}$$

for some $C_1 > 0$ and $q > 1$.
Then estimator $\widehat{g}_\lambda$ defined in (2.30) satisfies, for n great enough :

$$\mathbb{E}R(\widehat{g}_\lambda) - R(g^\star) \leq C\left(\frac{c(\lambda)}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}},$$

where $C = C(C_1, C_2, \kappa, \kappa_0, \rho, q) > 0$.

[8] This theorem requires the unicity of the Bayes $g^\star$. Such a restriction can be avoided using a more sophisticated geometry in Section 2.2.3.

This upper bound generalizes the result presented in Tsybakov [2004] or Koltchinskii [2006] to the indirect framework. Theorem 1 provides rates of convergence :

$$\left(c(\lambda)/\sqrt{n}\right)^{2\kappa/2\kappa+\rho-1}.$$

In the noise free case, with standard ERM estimators, Tsybakov [2004] or Koltchinskii [2006] obtain fast rates $n^{-\kappa/2\kappa+\rho-1}$. In the presence of contaminated inputs, rates are slower since $c(\lambda)\to+\infty$ as $n\to+\infty$. Hence, Theorem 1 shows that the Lipschitz constant introduced in Definition 2 is seminal in our problem. It gives the price to pay for the inverse problem in the statement of fast rates.

The behavior of the Lipschitz constant $c(\lambda)$ depend on the difficulty of the inverse problem through the degree of ill-posedness of operator A. This dissertation proposes to deal with mildly ill-posed inverse problems. In this case, $c(\lambda)$ depends polynomially on λ. Severely ill-posed inverse problems could be considered in future works, where in this case fast rates are prohibited.

2.2.2 Applications

Now, we turn out into several applications of the general setting of Theorem 5. In the problem of inverse statistical learning, we deal with a general compact operator A. The main consequence is that A^*A, where A^* denotes the adjoint of operator A, is not continuously inversible. To overcome this difficulty, several regularization methods have been proposed over years, such as Tikhonov type regularizations, recursive procedures in Hilbert space, or projection (or spectral cut-off) methods. These regularization schemes are often associated with the singular values decomposition (SVD) of A. Indeed, by compactness of A and spectral theorem (see Halmos [1963]), there exists an orthonormal basis $(\phi_k)_{k\in\mathbb{N}^*}$ of $L^2(\mathbb{R}^d)$ with associated eigenvalues $(b_k^2)_{k\in\mathbb{N}^*}$. The singular value decomposition of A is written :

$$A\phi_k = b_k\varphi_k \text{ and } A^\star\varphi_k = b_k\phi_k,\ k\in\mathbb{N}^*, \tag{2.33}$$

where $\varphi_k(\cdot)$ is the normalized version of $A\phi_k$ for any $k\in\mathbb{N}^*$.

In the problem of multiclass classification with indirect observations, we observe corrupted inputs Z with density Af, where f in the density of the direct input X. Then, if $\mathcal{Y}=\{0,\dots,M\}$, with (2.33) and by linearity of the operator, we may also write :

$$\forall y\in\mathcal{Y}, f_y=\sum_{k\in\mathbb{N}^*} b_k^{-1}\langle Af_y,\varphi_k\rangle\phi_k.$$

Then, the family of projection estimators $\hat{f}_y(\cdot)$ of each $f_y(\cdot)$, $y\in\mathcal{Y}$ has the form :

$$\hat{f}_y(\cdot)=\sum_{k=1}^{N}\hat{\theta}_k^y\phi_k(\cdot) \text{ where } \hat{\theta}_k^y=b_k^{-1}\frac{1}{n_y}\sum_{i=1}^{n_y}\varphi_k(Z_i^y), \tag{2.34}$$

whereas $N\geq 1$ is the regularization parameter. This is a particular case of representation (2.26) with :

$$k_N(x,z)=\sum_{k=1}^{N} b_k^{-1}\varphi_k(z)\phi_k(x).$$

Theorem 6. *Suppose $\{\ell(g)-\ell(g^\star), g\in\mathcal{G}\}$ is Bernstein class with respect to μ with parameter $\kappa\geq 1$. Suppose $0<\rho<1$ exists such that (2.31) holds for some $C_2>0$. Suppose there exists $\beta\in\mathbb{R}_+$ such that :*

$$b_k\sim k^{-\beta}\text{as } k\to+\infty.$$

Then, for n great enough, the minimizer of (2.27) with $k_N(\cdot,\cdot)$ defined above satisfies :

$$\mathbb{E}R(\widehat{g}_N)-R(g^\star)\leq Cn^{-\frac{\kappa\gamma}{\gamma(2\kappa+\rho-1)+(2\kappa-1)\beta}},$$

where $C=C(\gamma,L,C_2,\rho,\kappa,\kappa_0)$ and N is chosen such that :

$$N=n^{\frac{2\kappa-1}{2\gamma(2\kappa+\rho-1)+2(2\kappa-1)\beta}},$$

and for any $y \in \mathcal{Y}$, f_y is a bounded density with respect to Lebesgue contained in $\Theta(\gamma, L)$, the ellipsoïd in the SVD basis defined as :

$$\Theta(\gamma, L) = \{f = \sum_{k\geq 1} \theta_k \phi_k \in L_2(\mathcal{X}) : \sum_{k\geq 1} \theta_k^2 k^{2\gamma+1} \leq L\}.$$

Theorem 6 highlights fast rates of convergence under standard complexity and margin assumptions over the class $\mathcal{G}$. Gathering with a noise assumption related with the spectrum of the compact operator A^*A, we lead to fast rates of convergence provided that the conditional densities are sufficiently smooth, where the smoothness is related with the SVD of operator A. Here again, if $\beta = 0$, rates of Theorem 6 coincides with previous fast rates in the iterature. The price to pay for the ill-posedness is summarized in this case by $(2\kappa - 1)\beta/\gamma$. This term shows a strong dependence between the degree of ill-posedness, the margin (or Bernstein) assumption and the regularity of the conditional densities.

Many other cases have be considered by applying the general methodology of this section. We can state similar results in multiclass classification with errors in variables, with possible anisotropic shape. We can also consider the non-exact case as in Lecué and Mendelson [2012], where excess risk bounds are replaced by non-exact oracle inequalities of the following form :

$$\mathbb{E}R(\widehat{g}) \leq (1+\epsilon)R(g^\star) + C\psi_n.$$

These results are compiled in [L4], [L10] and [L16]. Other problems could be investigated such as learning principal curves (see Biau and Fisher [2012]), quantile estimation (Hall and Lahiri [2008] or Dattner, Reiß, and Trabs [2013]), level set estimation, or anomaly detection. The last risk bounds of this chapter are dedicated to the problem of clustering with errors in variables. This particular case has been the starting point of many developments that will be presented in the rest of Chapter 2-3.

2.2.3 Noisy clustering

One of the most popular issue in data mining or machine learning is to learn clusters from a big cloud of data. This problem is known as clustering. It has received many attention in the last decades. In this paragraph, we apply the general methodology of this section to the framework of noisy clustering. To frame the problem of noisy clustering into the general study of this chapter, we first introduce the following notations. Let X a $\mathbb{R}^d$-random variable with unknown density f with respect to the Lebesgue measure, such that $X \leq 1$ almost-surely. For some known integer $k \geq 1$, we introduce a set of codebooks $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$, and the standard k-means loss function $\ell(\mathbf{c}, X) := \min_{j=1,\ldots,k} |X - c_j|_2^2$, where $|\cdot|_2$ stands for the Euclidean norm in $\mathbb{R}^d$. The corresponding clustering risk of a codebook $\mathbf{c}$ is given by :

$$R(\mathbf{c}) := \mathbb{E}\ell(\mathbf{c}, X) = \int_{\mathbb{R}^d} \ell(\mathbf{c}, x) f(x) dx. \tag{2.35}$$

Given (2.35), we measure the performance of the latter codebook $\mathbf{c}$ in terms of excess risk, defined as :

$$R(\mathbf{c}) - R(\mathbf{c}^\star), \tag{2.36}$$

where $\mathbf{c}^\star \in \arg\min R(\mathbf{c})$ is called an *oracle*. The oracle set is denoted by $\mathcal{M}$ and we assume in the sequel that the number $|\mathcal{M}|$ of oracles is finite. This assumption is satisfied in the context of Pollard's regularity assumptions (see Pollard [1982]), i.e. when f has a continuous density (w.r.t. the Lebesgue measure) such that the Hessian matrix of $\mathbf{c} \mapsto R(\mathbf{c})$ is positive definite (see assumption **PRC** below). In the direct case, the problem of minimizing (2.36) has been investigated in a variety of areas. For a given number of clusters $k \geq 1$, the most popular technique is the k-means procedure. It consists in partitioning the dataset $X_1, \ldots, X_n$ into k clusters by minimizing the empirical risk :

$$\widehat{R}(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,k} |X_i - c_j|_2^2,$$

where $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$ is a set of centers. A cluster is associated to each observation by giving its nearest center c_j, $j = 1, \ldots, k$. The k-means clustering minimization has been widely studied in the literature. Since the early work of Pollard (Pollard [1981],Pollard [1982]), consistency and rates of convergence have been considered by many authors. Biau, Devroye, and Lugosi [2008] suggest rates of convergence of the form $\mathcal{O}(1/\sqrt{n})$ whereas Bartlett, Linder, and Lugosi [1998] propose a complete minimax study. More recently, Levrard [2013] states fast rates of the form $\mathcal{O}(1/n)$ under Pollard's regularity assumptions. It improves a previous result of Antos, Györfi, and György [2005].

In this section, we study the problem of clustering where we have at our disposal a corrupted sample $Z_i = X_i + \epsilon_i$, $i = 1, \dots, n$ where the ϵ_i's are i.i.d. with density η. By applying the same paths as in Section 2.1, we obtain a collection of noisy k-means minimizers :

$$\hat{\mathbf{c}}_h := \arg\min_{\mathbf{c}\in\mathcal{C}} \widehat{R}_h(\mathbf{c}), \quad h > 0, \tag{2.37}$$

where $\widehat{R}_h(\mathbf{c})$ is the deconvolution empirical risk for our problem. This quantity is based on the deconvolution kernel density estimator $\hat{f}_h(\cdot)$ defined in (2.3) as follows :

$$\widehat{R}_h(\mathbf{c}) = \int_{\mathcal{B}(0,1)} \ell(\mathbf{c}, x)\hat{f}_h(x)dx = \frac{1}{n}\sum_{i=1}^n \ell_h(\mathbf{c}, Z_i), \tag{2.38}$$

where $\ell_h(\mathbf{c}, Z)$ is the following convolution product :

$$\ell_h(\mathbf{c}, Z) := \left[\widetilde{\mathcal{K}}_h * (\ell(\mathbf{c},\cdot)\mathbf{1}_{\mathcal{B}(0,1)}(\cdot))\right](Z) = \int_{\mathcal{B}(0,1)} \widetilde{\mathcal{K}}_h(Z-x)\ell(\mathbf{c}, x)dx, \quad \mathbf{c} = (c_1, \dots, c_k) \in \mathcal{C},$$

with $\widetilde{\mathcal{K}}_h(\cdot)$ a deconvolution kernel and $\mathcal{C} := \left\{\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk} \,:\, c_j \in \mathcal{B}(0,1),\ j = 1, \dots, k\right\}$ is the set of possible centers. Remark that the restriction to the closed unit ball $\mathcal{B}(0,1)$ appears only for technicalities, since any compact set can be used.

To investigate the generalization ability of the family (2.37), we will use a localization technique inspired from Blanchard, Bousquet, and Massart [2008] (see also Levrard [2013]). As a first step, we derive in Theorem 7 fast rates of convergence, for a well-chosen non-adaptive [9] bandwidth parameter $h \in \mathbb{R}_+^d$.

In order to get satisfying upper bounds, we introduce the following regularity assumptions on the source distribution P.

Pollard's Regularity Condition (PRC) : The distribution P satisfies the following two conditions :

1. P has a continuous density f with respect to the Lebesgue measure on $\mathbb{R}^d$,
2. The Hessian matrix of $\mathbf{c} \longmapsto R(\mathbf{c})$ is positive definite for all optimal vector of clusters $\mathbf{c}^\star$.

It is easy to see that using the compactness of $\mathcal{B}(0, M)$, $\|X\|_\infty \leq M$ and **(PRC)** ensure that there exists only a finite number of optimal clusters $\mathbf{c}^\star \in \mathcal{M}$. This number is denoted by $|\mathcal{M}|$ in the rest of this section. Moreover, Pollard's condition ensures a margin assumption as in Section 2.1 thanks to the following lemma due to Antos et al. [5].

Lemma 4 (Antos et al. [5]). *Suppose $\|X\|_\infty \leq M$ and **(PRC)** holds. Then, for any $\mathbf{c} \in \mathcal{B}(0, M)$:*

$$\|\ell(\mathbf{c},\cdot) - \ell(\mathbf{c}^\star(\boldsymbol{c}),\cdot)\|_{L_2([0,1])}^2 \leq C_1\|\boldsymbol{c} - \boldsymbol{c}^\star(\boldsymbol{c})\|_2^2 \leq C_1C_2\left(R(\boldsymbol{c}) - R(\boldsymbol{c}^\star(\boldsymbol{c}))\right),$$

where $c^\star(\boldsymbol{c}) \in \arg\min_{\boldsymbol{c}^\star} \|\boldsymbol{c} - \boldsymbol{c}^\star\|_2$ and $\|\cdot\|_2$ stands for the Euclidean norm in the space of codebooks $\mathbb{R}^{dk}$.

Lemma 4 ensures a Bernstein assumption for the class $\mathcal{F} = \{\ell(\mathbf{c},\cdot) - \ell(\mathbf{c}^\star(\mathbf{c}),\cdot), \mathbf{c} \in \mathcal{B}(0, M)\}$ (see Definition 3). It is useful to derive fast rates of convergence. Recently, Levrard [2013] has pointed out sufficient conditions to have **(PRC)** when the source distribution P is well concentrated around its optimal clusters. From this point of view, Pollard's regularity conditions can be related to the margin assumption in binary classification.

[9] Chapter 3 investigates the construction of data-driven bandwidth parameters

Moreover, as in Section 2.1, we also need a noise assumption, which gives the behaviour of the characteristic function of the noise distribution. In the sequel, we use the following weaker assumption :

Noise Assumption NA(ρ,β). There exists some vector $\beta=(\beta_1,\dots,\beta_d)\in(0,\infty)^d$ and some positive constant ρ such that $\forall t\in\mathbb{R}^d$:

$$|\mathcal{F}[\eta](t)|\geq\rho\prod_{v=1}^{d}\left(\frac{t_v^2+1}{2}\right)^{\text{-}\beta_v/2}.$$

$\mathbf{NA}(\rho,\beta)$ deals with a lower bound on the behaviour of the characteristic function of the noise density η. This lower bound is a sufficient condition to get excess risk bounds. However, as mentioned above, to study the optimality in the minimax sense, we need an upper bound of the same order for the characteristic function.

Eventually, in this paragraph, we also extend the previous results to an anisotropic behaviour of the density f. It allows to consider more general classes of functions where the regularity depends on the direction. This regularity will be expressed in terms of anisotropic Hölder spaces.

Definition 5. *For some $s=(s_1,\dots,s_d)\in\mathbb{R}_d^+$, $L>0$, we say that f belongs to the anisotropic Hölder space $\Sigma(s,L)$ if the following holds :*

— *the function f admits derivatives with respect to x_j up to order $\lfloor s_j\rfloor$, where $\lfloor s_j\rfloor$ denotes the largest integer strictly less than s_j.*

— *$\forall j=1,\dots,d$, $\forall x\in\mathbb{R}^d$, $\forall x_j'\in\mathbb{R}$, the following Lipschitz condition holds :*

$$\left|\frac{\partial^{\lfloor s_j\rfloor}}{(\partial x_j)^{\lfloor s_j\rfloor}}f(x_1,\dots,x_{j-1},x_j',x_{j+1},\dots,x_d)-\frac{\partial^{\lfloor s_j\rfloor}}{(\partial x_j)^{\lfloor s_j\rfloor}}f(x)\right|\leq L|x_j'-x_j|^{s_j-\lfloor s_j\rfloor}.$$

If a function f belongs to the anisotropic Hölder space $\Sigma(s,L)$, f has an Hölder regularity s_j in each direction $j=1,\dots,d$. As a result, it can be well-approximated pointwise using a d-dimensional Taylor formula.

For this purpose, we require the following assumption on the kernel $\mathcal{K}$ which appears in $\widetilde{\mathcal{K}}_h$ (see (2.3)). This property looks like the previous assumption in the minimax theory of Section 2.1 with some minor changes due to the anisotropic framework.

Definition 6. *A kernel $\mathcal{K}$ is of order $m=(m_1,\dots,m_d)\in\mathbb{N}^d$ if and only if :*

— $\int_{\mathbb{R}^d}\mathcal{K}(x)dx=1$

— $\int_{\mathbb{R}^d}\mathcal{K}(x)x_j^k dx=0$, $\forall k\leq m_j$, $\forall j\in\{1,\dots,d\}$.

— $\int_{\mathbb{R}^d}|\mathcal{K}(x)||x_j|^{m_j}dx<K_2$, $\forall j\in\{1,\dots,d\}$.

We are now ready to state the main result of this paragraph.

Theorem 7. *Assume that $\mathbf{NA}(\rho,\beta)$ is satisfied for some $\beta\in(1/2,\infty)^d$, $\rho>0$ and **(PRC)** holds. Suppose $\eta_\infty:=\|\eta\|_\infty<\infty$ and $f\in\Sigma(s,L)$ with $L>0$ and $s\in\mathbb{R}_+^d$. Denote by $\hat{\mathbf{c}}_{\bar{h}}$ a solution of (2.37) with :*

$$\forall j=1,\dots,d,\overline{h}_j=n^{-1/(2s_j(1+\sum_{u=1}^d\beta_j/s_j))},$$

where $\bar{\beta}=\sum_{v=1}^d\beta_v$. Then, there exists a universal constant C_1 depending on $w,L,d,s,\beta,\rho,k,\eta_\infty$ and $|\mathcal{M}|$, and an integer $n_0\in\mathbb{N}^*$ such that for any $\mathbf{c}^\star\in\mathcal{M}$ and any $n\geq n_0$:

$$\mathbb{E}R(\hat{\mathbf{c}}_{\bar{h}},\mathbf{c}^\star)\leq C_1 n^{-1/(1+\sum_{j=1}^d\beta_j/s_j)}.$$

The proof is an application of a localization approach in the spirit of Massart [2007], applied to the noisy set-up [10]. As in Section 2.1, the bias variance decomposition (3.6) allows us to control the excess

[10] The main ingredient of the proof of Theorem 7 differs from Theorem 5 above. In the particular case of a finite dimensional space $\mathcal{G}$, proof's techniques from Theorem 5 are not optimal. Then, applied directly in noisy clustering, this result gives an extra $\sqrt{\log\log(n)}$ term in the RHS. This drawback comes from the localization scheme used in Theorem 5, which consists in using iteratively a Talagrand's type inequality. In the finite dimensional setting, we pay the number of iterations (i.e. an extra $\sqrt{\log\log(n)}$) in the upper bound.

risk. More precisely, the variance can be controlled by mixing empirical process as argued in Blanchard, Bousquet, and Massart [2008], gathering with the noise assumption $\mathbf{NA}(\rho,\beta)$. The bias term is bounded using both the smoothness of f and the margin assumption.

Rates of convergence of Theorem 2 are fast rates when $\bar{\beta} < \gamma$. It generalizes the result of Levrard [2013] to the errors-in-variables case since we can see coarsely that rates to the order $\mathcal{O}(1/n)$ are reached when $\epsilon = 0$. Here, the price to pay for the inverse problem is the quantity $\sum_{i=1}^{d} \beta_i$, related to the tail behavior of the characteristic function of the noise distribution η in $\mathbf{NA}(\rho,\beta)$.

2.3 A new algorithm for noisy clustering [L10]

When we consider direct data $X_1, \ldots, X_n$, we are interested in the minimization of the empirical risk $\sum_{i=1}^{n} \min_{j=1,\ldots,k} |X_i - c_j|_2^2$. In this respect, the basic iterative procedure of k-means was proposed by Lloyd in a seminal work (Lloyd [1982], first published in 1957 in a technical note of Bell laboratories). The procedure calculates, from an initialization of k centers, the associated Voronoï cells and updates the centers with the means of the data on each Voronoï cell. Bubeck [2002] has shown that it corresponds exactly to a step of a Newton optimization. The k-means with Lloyd algorithm is considered as a staple in the study of clustering methods. The time complexity is approximately linear, and appears as a good algorithm for clustering spherical well-separated classes, such as a mixture of gaussian vectors.

However, in many real-life situations, direct data are not available and measurement errors may occur. In social science, many data are collected by human pollster, with a possible contamination in the survey process. In medical trials, where chemical or physical measurements are treated, the diagnostic is affected by many nuisance parameters, such as the measuring accuracy of the considered machine, gathering with a possible operator bias due to the human practitionner. Same kinds of phenomenon occur in astronomy or econometrics (see Meister [2009]). However, to the best of our knowledge, these considerations are not taken into account in the clustering task. The main implicit argument is that these errors have zero mean and could be neglected at the first glance. In this section we design a novel algorithm to perform clustering over contaminated datasets. We show that it can significantly improve the expected performances of a standard clustering algorithm which neglect this additional source of randomness.

2.3.1 First order conditions

When considering indirect data $Z_i = X_i + \epsilon_i$, $i = 1, \ldots, n$, a deconvolution empirical risk is defined as :

$$\frac{1}{n}\sum_{i=1}^{n} \ell_h(\mathbf{c}, Z_i) = \int_{[0,M]^d} \min_{j=1,\ldots,k} |x - c_j|_2^2 \hat{f}_h(x)dx. \tag{2.39}$$

Reasonably, a noisy clustering algorithm could be adapted, following the direct case and the construction of the standard k-means. The following theorem gives the first order conditions to minimize the deconvolution empirical risk (2.39). In the sequel, $\nabla F(x)$ denotes the gradient of a function $F : \mathbb{R}^{dk} \to \mathbb{R}$ at point $x \in \mathbb{R}^{dk}$.

Theorem 8. *Suppose assumptions of Theorem 7 are satisfied. Then, for any $h > 0$:*

$$\nabla \sum_{i=1}^{n} \ell_h(\bar{\mathbf{c}}, Z_i) = 0_{\mathbb{R}^{dk}},$$

where :

$$\bar{\mathbf{c}}_{u,j} = \frac{\sum_{i=1}^{n} \int_{V_j} x_u \widetilde{\mathcal{K}}_h(Z_i - x)dx}{\sum_{i=1}^{n} \int_{V_j} \widetilde{\mathcal{K}}_h(Z_i - x)dx}, \forall u \in \{1, \ldots, d\}, \forall j \in \{1, \ldots, k\}, \tag{2.40}$$

where $\bar{\mathbf{c}}_{u,j}$ stands for the u-th coordinates of the j-th centers, whereas V_j is the Voronoï cell of $\bar{\mathbf{c}}$ with center j :

$$V_j = \{x \in \mathbb{R}^d : \min_{j'=1,\ldots,k} |x - c_{j'}|_2 = |x - c_j|_2\}.$$

```
1. Initialize the centers $\mathbf{c}^{(0)} = (c_1^{(0)}, \ldots, c_k^{(0)}) \in \mathbb{R}^{dk}$
2. Estimation step :
   (a) Compute the deconvoluting Kernel $\mathcal{K}_\eta$ and its FFT $\mathcal{F}(\mathcal{K}_\eta)$.
   (b) Build a histogram of 2-d grid using linear binning rule and compute its FFT : $\mathcal{F}(\hat{f}_Z)$.
   (c) Compute : $\mathcal{F}(\hat{f}) = \mathcal{F}(\mathcal{K}_\eta)\mathcal{F}(\hat{f}_Z)$.
   (d) Compute the Inverse FFT of $\mathcal{F}(\hat{f})$ to obtain the density estimated of X : $\hat{f} = \mathcal{F}^{-1}(\mathcal{F}(\hat{f}))$.
3. Repeat until convergence :
   (a) Assign data points to closest clusters in order to compute the Voronoi diagram.
   (b) Re-adjust the center of clusters with equation (2.41).
4. Compute the final partition by assigning data points to the final closest clusters $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_k)$.
```

Figure 2.1 : The algorithm of Noisy k-means.

The proof is based on the calculation of the directional derivatives of the deconvolution empirical risk (2.39). It is easy to see that a similar result can be shown with the k-means. Indeed, a necessary condition in the direct minimization problem is as follows :

$$\mathbf{c}_{u,j} = \frac{\sum_{i=1}^n \int_{V_j} x_u \delta_{X_i} dx}{\sum_{i=1}^n \int_{V_j} \delta_{X_i} dx}, \; \forall u \in \{1, \ldots, d\}, \forall j \in \{1, \ldots, k\},$$

where δ_{X_i} is the Dirac function at point X_i. Theorem 8 proposes a same kind of condition in the errors-in-variable case replacing the Dirac function by a deconvolution kernel. We can also perceive that by switching the integral with the sum in equation (2.40), the first order conditions on $\mathbf{c}$ can be rewritten as follows :

$$\bar{\mathbf{c}}_{u,j} = \frac{\int_{V_j} x_u \hat{f}_h(x) dx}{\int_{V_j} \hat{f}_h(x) dx}, \; \forall u \in \{1, \ldots, d\}, \; \forall j \in \{1, \ldots, k\}, \tag{2.41}$$

where $\hat{f}_h(x) = 1/n \sum_{i=1}^n \mathcal{K}_\eta \left(Z_i - x/h \right)/h$ is the kernel deconvolution estimator of the density f. This property is at the core of the algorithm presented in Figure 2.1.

Eventually, the expression of $\bar{\mathbf{c}}$ in Theorem 8 can lead to a kernelized version of the algorithm in the noiseless case. Indeed, we can replace the deconvolution kernel by a standard kernel function (such as the indicator function) with a sufficiently small bandwidth. This idea has been already presented in Section 2.1 where optimality in the minimax sense is proved in discriminant analysis (see Corollary 1).

2.3.2 Experimental validation

Evaluation of clustering algorithms is not an easy task (see von Luxburg, Williamson, and Guyon [2009]). In supervised classification, cross-validation techniques are standard to evaluate learning algorithms such as classifiers. The principle is to divide the sample into V subsets, the first $V-1$ are used for training the considered classifiers whereas the last one is used for testing these classifiers. Unfortunately, in an unsupervised framework - such as clustering - the performances of new algorithms depend on what one is trying to do. In this section, we propose two experimental settings to illustrate the efficiency of noisy k-means with different criteria based on clustering or Euclidean distance.

These experimental settings are based on simulations of gaussian mixtures with additive random noise. We want to emphasize that this additional source of randomness does not have to be neglected for both clustering or quantization. For this purpose, we compare noisy k-means algorithm (based on a deconvolution step) with standard k-means (a direct algorithm) using Lloyd algorithm, where the random initialization is common for both methods. It allows us to reduce the dependence to the initialization of the measure of performances, due to the non-convexity of the considered problem (see Bubeck [2002]). Eventually, we will see that this comparison depends on several parameters in our models, such as the level of noise ϵ, the type of noise (Laplace or Gaussian) and the number $k = 2$ or $k = 4$ of Gaussian mixtures.

Experimental setting

We consider two different experiments $j = 1, 2$ based on i.i.d. noisy samples $\mathcal{D}_n^{(j)} = \{Z_1^{(j)}, \ldots, Z_n^{(j)}\}$ where :

$$Z_i^{(j)} = X_i^{(j)} + \epsilon_i(u),\ i = 1, \ldots, n, \quad \mathbf{Modj}(\mathbb{L}, u), \tag{2.42}$$

where $(X_i^{(j)})_{i=1}^n$ are i.i.d. with density $f^{(j)}$ where :

— $f^{(1)} = 1/2 f_{\mathcal{N}(0_2, I_2)} + 1/2 f_{\mathcal{N}((5,0)^T, I_2)}$, whereas

— $f^{(2)} = 1/4 f_{\mathcal{N}(0_2, I_2)} + 1/4 f_{\mathcal{N}((5,0)^T, I_2)} + 1/4 f_{\mathcal{N}((0,5)^T, I_2)} + 1/4 f_{\mathcal{N}((5,5)^T, I_2)}$.

Moreover, $(\epsilon_i(u))_{i=1}^n$ are i.i.d. with law $\mathbb{L}$ with zero mean $(0,0)^T$ and covariance matrix $\Sigma(u) = \begin{pmatrix} 1 & 0 \\ 0 & u \end{pmatrix}$ for $u \in \{1, \ldots, 10\}$. We consider two cases for $\mathbb{L}$, namely a two-dimensional Laplace ($\mathcal{L}$) or Gaussian ($\mathcal{N}$) noise.

For each experiment $j = 1, 2$, we propose to compare the performances of Noisy k-means with respect to k-means by computing three different criteria. Given a noisy sample $Z_i = X_i + \epsilon_i$, $i = 1, \ldots, n$, we compute the clusterring error according to :

$$\mathcal{I}_n(\hat{\mathbf{c}}) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}(Y_i \neq f_{\hat{\mathbf{c}}}(X_i)),\ \forall \hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_k) \in \mathbb{R}^{dk}, \tag{2.43}$$

where $f_{\hat{\mathbf{c}}}(x) = \arg\min_{j=1,\ldots,k} |x - \hat{\mathbf{c}}_j|_2^2$ and $Y_i \in \{1, 2\}$ for $j = 1$ (resp. $Y_i \in \{1, 2, 3, 4\}$ for $j = 2$) corresponds to the mixture of the point X_i. We also compute the quantization error $\mathcal{Q}_n(\hat{\mathbf{c}})$ defined as :

$$\mathcal{Q}_n(\hat{\mathbf{c}}) := \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,k} |X_i - \hat{c}_j|_2^2,\ \forall \hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_k) \in \mathbb{R}^{dk}. \tag{2.44}$$

Eventually, from an estimation point of view, we can also compute the ℓ_2-estimation error of $\hat{\mathbf{c}}$ given by :

$$\|\hat{\mathbf{c}} - \mathbf{c}^\star\|_2 := \sqrt{\sum_{j=1}^k |\hat{c}_j - c_j^\star|_2^2},\ \forall \hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_k) \in \mathbb{R}^{dk}, \tag{2.45}$$

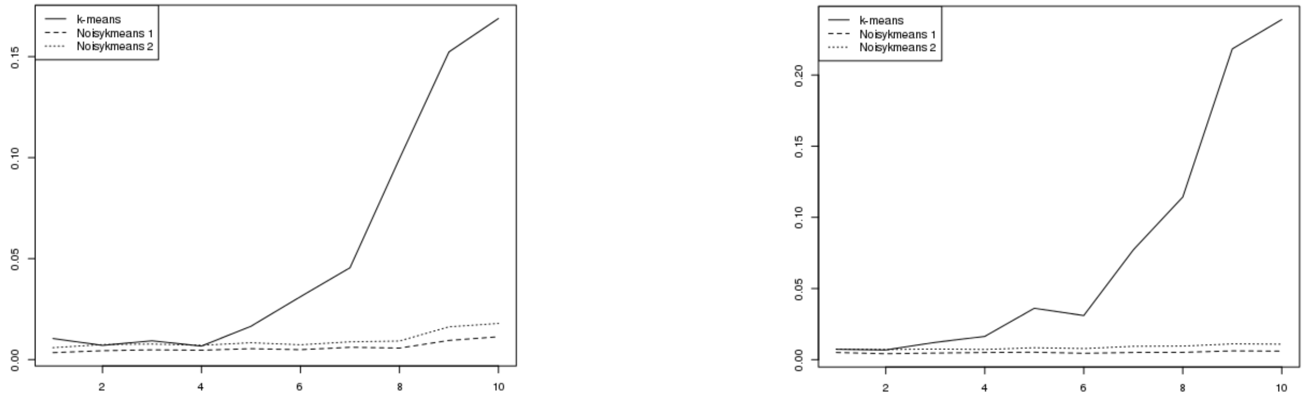
where $(c_1^\star, c_2^\star) = (0, 0, 5, 0)$ for $\mathbf{Mod1}(\mathbb{L}, u)$ (resp. $(c_1^\star, c_2^\star, c_3^\star, c_4^\star) = (0, 0, 5, 0, 0, 5, 5, 5)$ for $\mathbf{Mod2}(\mathbb{L}, u)$).

For each criterion, we study the behaviour of the Lloyd algorithm (standard k-means) with two different noisy k-means, corresponding to two different choice of bandwidths h in the estimation step (see Figure 2.1). For a grid $h \subseteq [0.1, 5]^2$ of 10×10 parameters, we compute $h_{\mathcal{I}}$ defined as the minimizer of (2.43) over the grid h whereas $h_{\mathcal{Q}}$ is the minimizer of (2.44). Then, we have three clustering algorithms denoted by $\hat{\mathbf{c}}$ for standard k-means using Lloyd algorithm, and $\{\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2\}$ for noisy k-means algorithms with the same initialization and with associated bandwidth $h_{\mathcal{I}}$ and $h_{\mathcal{Q}}$ defined above. It is important to stress that choice of bandwidth $h_{\mathcal{I}}$ and $h_{\mathcal{Q}}$ are not possible in practice. Hence, an adaptive procedure to choose the bandwidth has to be performed, as in standard nonparametric problems. This is out of the scope of the present chapter where we propose to compare k-means with Noisy k-means with fixed bandwidths $h_{\mathcal{Q}}$ and $h_{\mathcal{I}}$. In the sequel, we illustrate the behaviour of these methods for each criterion and each experiment.

Results of the first experiment

In the first experiment, we run 100 realizations of training set $\{Z_1^{(1)}, \ldots, Z_n^{(1)}\}$ from (3.36) with $n = 200$. At each realization, we run Lloyd algorithm and noisy k-means with the same random initialization.

Clustering risk Figure 2.2 (a)-(b) illustrates the evolution of the clustering risk (2.43) of $\{\hat{\mathbf{c}}, \hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2\}$ when $u \in \{1, \ldots, 10\}$ (horizontal axe) in $\mathbf{Mod1}(\mathbb{L}, u)$.

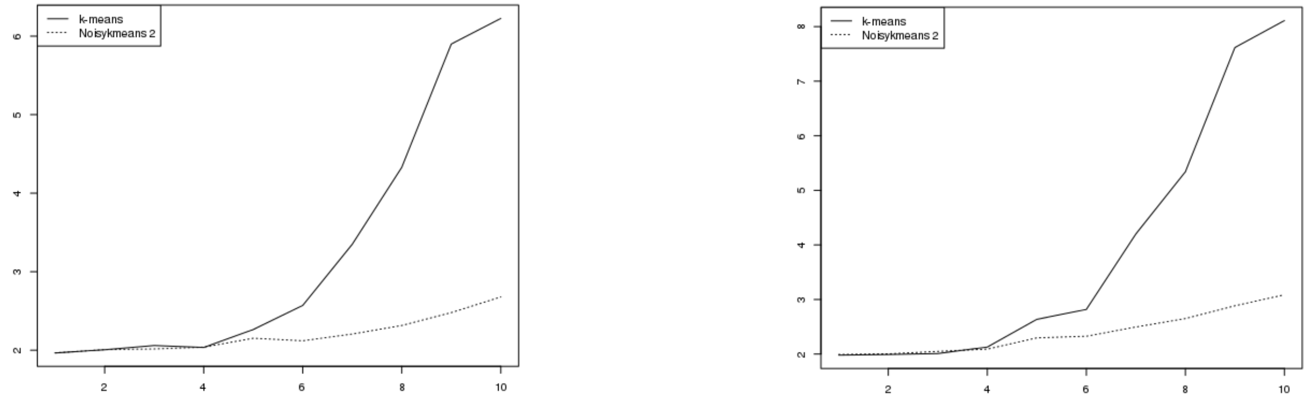


(a) Laplace error (b) Gaussian error

Figure 2.2 : Clustering risk averaged over 100 replications from $\mathbf{Mod1}(\mathbb{L}, u)$ with $n = 200$.

When $u \leq 4$, the results are comparable and Noisy k-means seems to slightly outperform standard k-means. However, when the level of noise in the vertical axe becomes higher (i.e. $u \geq 5$), k-means with Lloyd algorithm shows a very bad behaviour. On the contrary, noisy k-means seems robust in these situations, for both Laplace and Gaussian noise.

Quantization risk Figure 2.3 (a)-(b) shows the behaviour of the quantization risk (2.44) of $\hat{\mathbf{c}}$ and $\hat{\mathbf{c}}_{\mathcal{Q}}$ when u increases.

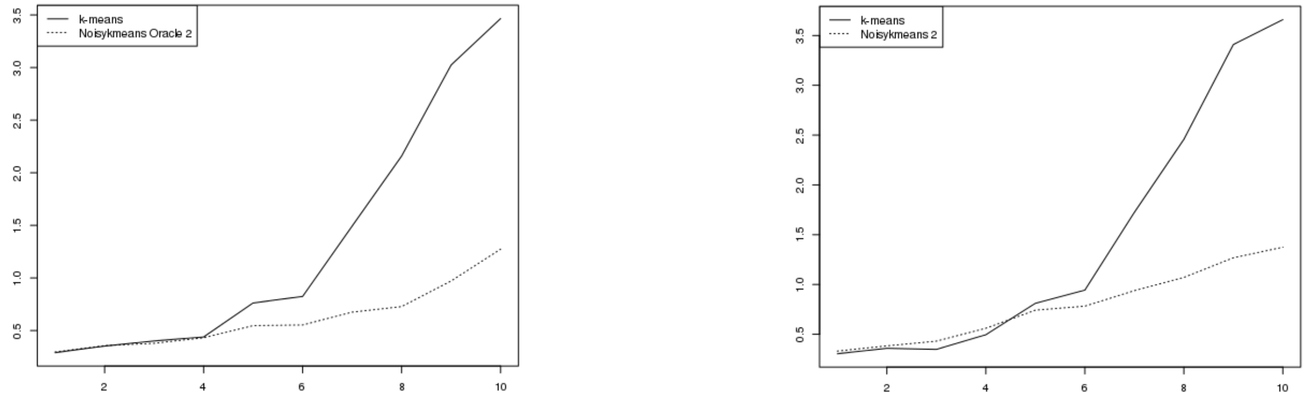


(a) Laplace error (b) Gaussian error

Figure 2.3 : Quantization risk averaged over 100 replications from $\mathbf{Mod1}(\mathbb{L}, u)$ with $n = 200$.

We omit $\hat{\mathbf{c}}_1$ because it shows bad performances when the variance u in $\mathbf{Mod1(\mathbb{L},u)}$ increases (see Table 2.1). This phenomenon can be explained as follows : $\hat{\mathbf{c}}_1$ is chosen to minimize the clustering risk (2.43). As a result, the proposed codebook $\hat{\mathbf{c}}_1$ is not necessarily a good quantizer, even if it gives good Voronoï cells for clustering the set of data. On the contrary, $\hat{\mathbf{c}}_2$ outperform standard k-means when the vertical variance increases. The quantization error behaves like the clustering risk above. Laplace and Gaussian noise highlight comparable results.

L2 risk In Figure 2.4 (a)-(b), the ℓ_2 risk (2.45) of $\hat{\mathbf{c}}$ and $\hat{\mathbf{c}}_2$ is proposed. In this case, we can see a more efficient robustness to the noise for Noisy k-means in comparison with standard k-means. However, in comparison with the two other criteria, the ℓ_2 risk of noisy k-means increases when the variance increases. This phenomenon is comparable for Laplace and Gaussian noise, with a slightly better robustness of noisy k-means in the Laplace case.



(a) Laplace error (b) Gaussian error

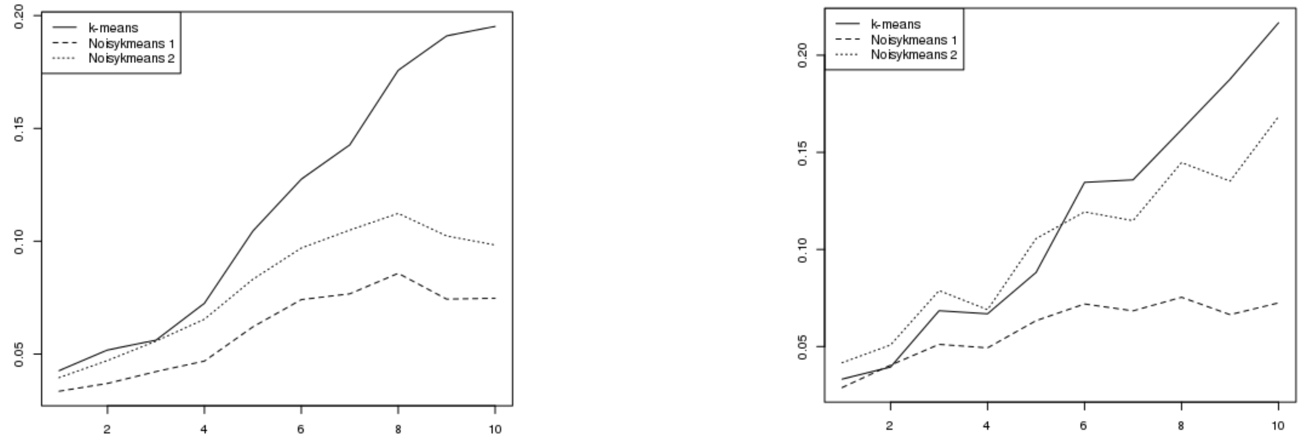
Figure 2.4 : ℓ_2-risk averaged over 100 replications from $\mathbf{Mod1}(\mathbb{L}, u)$ with $n = 200$.

Conclusion of the first experiment The first experiment shows very well the lack of efficiency of the standard k-means when we deal with errors in variables. When the variance of the noise ϵ increases, the performances of the k-means are deteriorated. On the contrary, the noisy k-means shows a good robustness to this additional source of noise for the considered criteria.

Result of the second experiment

In the second experiment, we run 100 realizations of training set $\{Z_1^{(2)}, \ldots, Z_n^{(2)}\}$ from (3.36) with $n = 200$. At each realization, we run Lloyd algorithm and Noisy k-means with the same random initialization.

Clustering risk Figure 2.5 (a)-(b) shows the evolution of the clustering risk (2.43) of $\{\hat{\mathbf{c}}, \hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2\}$ when $u \in \{1, \ldots, 10\}$ in **Mod2(𝕃,u)** is proposed.

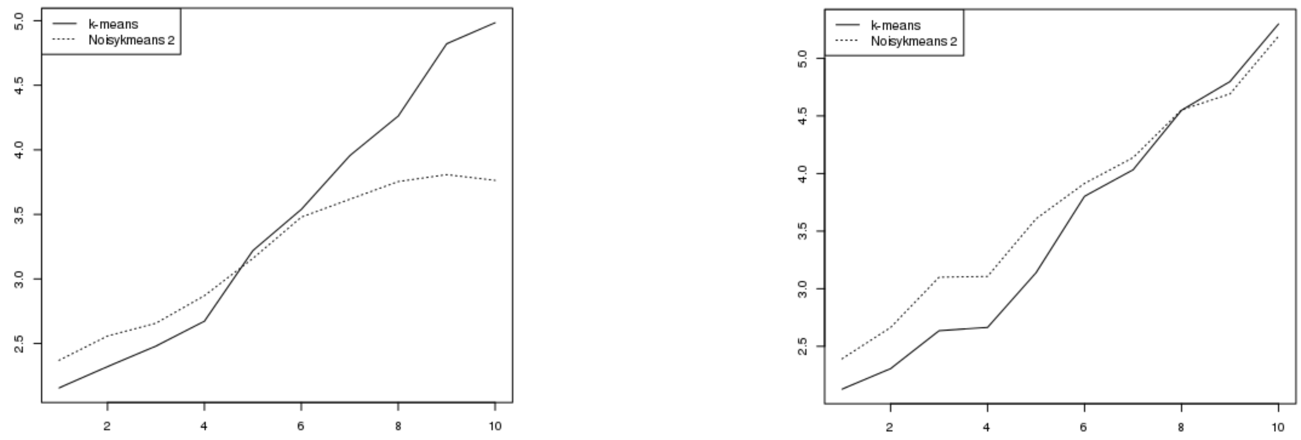


(a) Laplace error (b) Gaussian error

Figure 2.5 : Clustering risk averaged over 100 replications from $\mathbf{Mod2}(\mathbb{L}, u)$ with $n = 200$.

Figure 2.5 shows a good resistance of noisy k-means $\hat{\mathbf{c}}_1$ in the presence of a mixture of four Gaussian with errors. When the level of noise is small, $\hat{\mathbf{c}}_1$ slightly outperforms k-means $\hat{\mathbf{c}}$ and when the level of noise becomes higher (i.e. $u \geq 5$), k-means with Lloyd algorithm shows a very bad behaviour. On the contrary, noisy k-means seems more robust in these situations. However, in the presence of a Gaussian noise, $\hat{\mathbf{c}}_2$ is comparable with $\hat{\mathbf{c}}$.

Quantization risk Figure 2.6 (a)-(b) shows the evolution of the quantization risk (2.43) of $\hat{\mathbf{c}}$ and $\hat{\mathbf{c}}_2$ when $u \in \{1, \ldots, 10\}$ in **Mod(2,𝕃)**. We omit $\hat{\mathbf{c}}_1$ for the same reason as in **Mod1(𝕃,u)**.

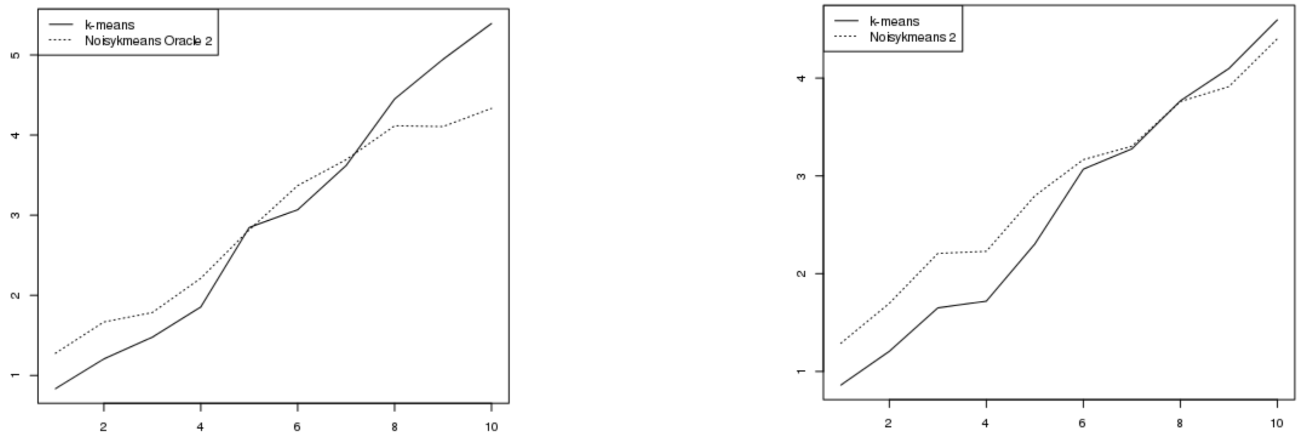


(a) Laplace error (b) Gaussian error

Figure 2.6 : Quantization risk averaged over 100 replications from $\mathbf{Mod2}(\mathbb{L}, u)$ with $n = 200$.

Here the evolution of the quantization risk depends strongly on the type of noise in $\mathbf{Mod2}(\mathbb{L}, u)$. When the noise is Laplace, $\hat{\mathbf{c}}_2$ outperforms standard k-means when the vertical variance $u \geq 5$, whereas for small variance, the results are comparable. On the contrary, when the additive noise is Gaussian, the problem seems intractable and Noisy k-means with $n = 200$ does not provide interesting results.

L2 risk Figure 2.7 (a)-(b) proposes the ℓ_2 risk (2.45) of $\hat{\mathbf{c}}$ and $\hat{\mathbf{c}}_2$ in $\mathbf{Mod2}(\mathbb{L}, u)$.



(a) Laplace error (b) Gaussian error

Figure 2.7 : ℓ_2-risk averaged over 100 replications from $\mathbf{Mod1}(\mathbb{L}, u)$ with $n = 200$.

The results are comparable with the Quantization risk and even worst : the Noisy k-means outperforms standard k-means for higher variance ($u \geq 8$).

Conclusion of the second experiment The performances of the k-means are deteriorated when the variance of ϵ increases in the second experiment. However, in this experiment, the problem of noisy clustering -or noisy quantization - seems more difficult. Indeed, Noisy k-means algorithms are not always significantly better than a standard k-means. In this experiment, the difficulty of the problem strongly depends on the type of noise (Gaussian or Laplace), which coincides with standard results in errors-in-variables models.

Conclusion of the experimental study

The results of this section show rather well the importance of the deconvolution step in the problem of clustering with errors-in-variables. In the presence of well-separated Gaussian mixtures with additive noise, standard k-means gives very bad performances when the variance of the noise increases. On the contrary, Noisy k-means is more robust to this additional source of randomness. In the particular case of the first experiment, noisy k-means significantly outperforms standard k-means. Unfortunately, when the mixture is more complicated (4 modes in the second experiment), the problem of noisy clustering seems more difficult. The performances of Noisy k-means are not as good as in the first experiment.

Conclusion of Chapter 2

This chapter furnishes the first few steps toward a general theory of statistical learning with a contaminated sample. Minimax fast rates of convergence are presented in the problem of discriminant analysis. Other risks bounds are proposed in a general setting, with possible applications. The end of the chapter focuses on the problem of clustering a noisy sample, where theoretical and practical issues are considered.

These chapter tries to study as precisely as possible the influence of the inverse problem over the statement of fast rates in classification. It appears that a key role is played by the spectrum of the operator, such as the behaviour of the characteristic function of ϵ in the deconvolution case. Many issues could be considered in the future. These questions are compiled at the end of the manuscript, where dozen of open problems are developed.

In the next chapter, we attack the problem of bandwidth selection, i.e. the construction of data-driven bandwidth selection methods in this context of statistical learning with a corrupted sample, or more generally, in kernel empirical risk minimization.

		$\mathcal{I}_n$		$\mathcal{Q}_n$		ℓ_2	
		Lap.	Gaus.	Lap.	Gaus.	Lap.	Gaus.
$\sigma = 1$	$\hat{\mathbf{c}}$	1.1	0.7	1.96	1.98	0.29	0.30
	$\hat{\mathbf{c}}_1$	0.3	0.5	2.28	3.39	0.62	1.02
	$\hat{\mathbf{c}}_2$	0.6	0.7	1.97	1.99	0.30	0.33
$\sigma = 2$	$\hat{\mathbf{c}}$	0.7	0.7	2.01	1.99	0.35	0.36
	$\hat{\mathbf{c}}_1$	0.4	0.4	2.42	2.86	0.77	0.94
	$\hat{\mathbf{c}}_2$	0.7	0.7	2.01	2	0.36	0.38
$\sigma = 3$	$\hat{\mathbf{c}}$	0.9	1.2	2.06	2.01	0.40	0.35
	$\hat{\mathbf{c}}_1$	0.5	0.5	2.35	2.83	0.71	0.90
	$\hat{\mathbf{c}}_2$	0.8	0.7	2.02	2.05	0.38	0.43
$\sigma = 4$	$\hat{\mathbf{c}}$	0.7	1.6	2.04	2.13	0.44	0.50
	$\hat{\mathbf{c}}_1$	0.5	0.5	2.35	3.65	0.79	1.28
	$\hat{\mathbf{c}}_2$	0.7	0.7	2.04	2.09	0.43	0.56
$\sigma = 5$	$\hat{\mathbf{c}}$	1.7	3.6	2.26	2.64	0.76	0.81
	$\hat{\mathbf{c}}_1$	0.5	0.5	2.72	3.90	1.05	1.45
	$\hat{\mathbf{c}}_2$	0.8	0.8	2.15	2.30	0.55	0.74
$\sigma = 6$	$\hat{\mathbf{c}}$	3.1	3.1	2.57	2.82	0.82	0.94
	$\hat{\mathbf{c}}_1$	0.5	0.5	2.70	3.87	1.08	1.62
	$\hat{\mathbf{c}}_2$	0.7	0.8	2.12	2.33	0.55	0.78
$\sigma = 7$	$\hat{\mathbf{c}}$	4.5	7.7	3.35	4.20	1.49	1.72
	$\hat{\mathbf{c}}_1$	0.6	0.5	2.96	3.93	1.30	1.61
	$\hat{\mathbf{c}}_2$	0.9	0.9	2.21	2.50	0.68	0.94
$\sigma = 8$	$\hat{\mathbf{c}}$	10.0	11.4	4.33	5.34	2.16	2.46
	$\hat{\mathbf{c}}_1$	0.6	0.5	3.29	4.51	1.46	1.82
	$\hat{\mathbf{c}}_2$	0.9	1	2.32	2.65	0.73	1.07
$\sigma = 9$	$\hat{\mathbf{c}}$	15.2	21.8	5.9	7.62	3.02	3.41
	$\hat{\mathbf{c}}_1$	1.0	0.6	3.69	5.29	1.67	2.14
	$\hat{\mathbf{c}}_2$	1.6	1.1	2.48	2.89	0.97	1.27
$\sigma = 10$	$\hat{\mathbf{c}}$	16.9	23.9	6.22	8.11	3.47	3.66
	$\hat{\mathbf{c}}_1$	1.1	0.6	3.85	5.27	1.84	2.21
	$\hat{\mathbf{c}}_2$	1.8	1.1	2.68	3.09	1.27	1.37

Table 2.1 : Results of the first experiments averaged over 100 replications. Quantities $\mathcal{I}_n$, $\mathcal{Q}_n$, ℓ_2 are defined in equations (2.43)-(2.45) whereas estimators $\hat{\mathbf{c}}$ (k-means with Lloyd), $\hat{\mathbf{c}}_1$ and $\hat{\mathbf{c}}_2$ (noisy k-means with two particular bandwidths) are defined in Section 2.3.2. The values of σ corresponds to the variance of the vertical direction of the additive noise ϵ, which is distributed as a Laplace or a Gaussian distribution).

		$\mathcal{I}_n$		$\mathcal{Q}_n$		ℓ_2	
		Lap.	Gaus.	Lap.	Gaus.	Lap.	Gaus.
$\sigma = 1$	$\hat{\mathbf{c}}$	4.3	3.3	2.16	2.13	0.83	0.86
	$\hat{\mathbf{c}}_1$	3.4	2.9	2.57	4.24	1.55	2.14
	$\hat{\mathbf{c}}_2$	4.0	4.2	2.37	2.39	1.28	1.29
$\sigma = 2$	$\hat{\mathbf{c}}$	5.2	3.9	2.32	2.31	1.21	1.21
	$\hat{\mathbf{c}}_1$	3.7	4.0	2.88	7.00	1.87	3.40
	$\hat{\mathbf{c}}_2$	4.7	5.1	2.56	2.66	1.67	1.70
$\sigma = 3$	$\hat{\mathbf{c}}$	5.6	6.8	2.48	2.64	1.48	1.65
	$\hat{\mathbf{c}}_1$	4.2	5.1	3.03	10.15	2.12	4.58
	$\hat{\mathbf{c}}_2$	5.6	7.9	2.66	3.10	1.79	2.21
$\sigma = 4$	$\hat{\mathbf{c}}$	7.3	6.7	2.67	2.66	1.85	1.72
	$\hat{\mathbf{c}}_1$	4.7	4.9	3.59	8.79	2.56	4.29
	$\hat{\mathbf{c}}_2$	6.5	6.9	2.87	3.11	2.21	2.23
$\sigma = 5$	$\hat{\mathbf{c}}$	10.5	8.8	3.22	3.14	2.85	2.30
	$\hat{\mathbf{c}}_1$	6.2	6.3	4.03	11.17	3.11	5.28
	$\hat{\mathbf{c}}_2$	8.3	10.6	3.16	3.61	2.82	2.80
$\sigma = 6$	$\hat{\mathbf{c}}$	12.8	13.5	3.54	3.80	3.07	3.07
	$\hat{\mathbf{c}}_1$	7.4	7.2	4.34	12.88	3.43	5.97
	$\hat{\mathbf{c}}_2$	9.7	11.9	3.48	3.91	3.37	3.17
$\sigma = 7$	$\hat{\mathbf{c}}$	14.3	13.6	3.95	4.03	3.62	3.28
	$\hat{\mathbf{c}}_1$	7.7	6.8	4.72	12.84	3.83	6.02
	$\hat{\mathbf{c}}_2$	10.5	11.5	3.62	4.14	3.69	3.30
$\sigma = 8$	$\hat{\mathbf{c}}$	17.6	16.2	4.26	4.55	4.45	3.77
	$\hat{\mathbf{c}}_1$	8.6	7.5	4.75	14.57	4.28	6.76
	$\hat{\mathbf{c}}_2$	11.2	14.5	3.75	4.55	4.12	3.76
$\sigma = 9$	$\hat{\mathbf{c}}$	19.1	18.8	4.82	4.80	4.95	4.10
	$\hat{\mathbf{c}}_1$	7.4	6.6	5.12	14.13	3.98	6.61
	$\hat{\mathbf{c}}_2$	10.2	13.5	3.81	4.69	4.11	3.91
$\sigma = 10$	$\hat{\mathbf{c}}$	19.5	21.7	4.98	5.30	5.39	4.60
	$\hat{\mathbf{c}}_1$	7.5	7.3	5.19	14.56	4.23	6.88
	$\hat{\mathbf{c}}_2$	9.8	16.8	3.76	5.19	4.33	4.40

Table 2.2 : Results of the second experiment averaged over 100 replications. Quantities $\mathcal{I}_n$, $\mathcal{Q}_n$, ℓ_2 are defined in equations (2.43)-(2.45) whereas estimators $\hat{\mathbf{c}}$ (k-means with Lloyd), $\hat{\mathbf{c}}_1$ and $\hat{\mathbf{c}}_2$ (noisy k-means with different bandwidths) are defined in Section 2.3.2. The values of σ corresponds to the variance of the vertical direction of the additive noise ϵ, which is distributed as a Laplace or a Gaussian distribution).

Chapitre 3

Bandwidth selection in kernel empirical risk minimization

The problem of bandwidth selection is fundamental in nonparametric statistics. In kernel density estimation, the starting point is a *bias-variance decomposition* according to :

$$\mathbb{E}|\hat{f}_h(x_0)-f(x_0)|^2 \leq |f_h(x_0)-f(x_0)|^2+\mathbb{E}|\hat{f}_h(x_0)-f_h(x_0)|^2 =: \mathrm{bias}(h)+\mathrm{var}(h), \tag{3.1}$$

where $\hat{f}_h(\cdot)=\sum_{i=1}^n \mathcal{K}_h(X_i-\cdot)/n$ is a kernel estimator of f based on a i.i.d. sample $(X_i)_{i=1}^n$ with mean $f_h(\cdot)=\mathbb{E}\hat{f}_h(\cdot)$. To state minimax rates of convergence, a deterministic choice of h trades off the bias term and the variance term in (3.1), and depends on unknown parameters, such as the smoothness index of the density f. Given a family $\{\hat{f}_h, h\in\mathcal{H}\}$, the problem of bandwidth selection is the data-driven selection of an estimator from this family which satisfies some *adaptive optimal properties* : the selected estimator reaches the minimax rate for any function in a vast range of regularities. In this case, the proposed bandwidth does not depend on the exact smoothness index of the target function but only on an upper bound.

In Chapter 2, non-adaptive fast rates of convergence have been derived for several ERM strategies based on a deconvolution kernel $\widetilde{\mathcal{K}}_h(\cdot)$. This kernel depends on some bandwidth parameter $h\in\mathbb{R}_+^d$ whose optimal calibration is critical. In particular, an appropriate choice of the bandwidth provides in Section 3.3 fast rates in noisy clustering thanks to the following bias-variance decomposition :

$$R(\hat{\mathbf{c}}_h,\mathbf{c}^\star)\leq (R-\widehat{R}_h)(\hat{\mathbf{c}}_h,\mathbf{c}^\star)\leq (R-R_h)(\hat{\mathbf{c}}_h,\mathbf{c}^\star)+(R_h-\widehat{R}_h)(\hat{\mathbf{c}}_h,\mathbf{c}^\star), \tag{3.2}$$

where exhaustive notations are presented in Section 3.1. We can perceive that - by and large - decomposition (3.1) and (3.2) have the same flavour.

One of the most popular method for choosing the bandwidth is suggested by Lepski, Mammen, and Spokoiny [1997] in a gaussian white noise model. It is based on the *Lepski's principle* (Lepski [1990]). The idea is to test several estimators (by comparison) for different values of the bandwidth. This work is at the origin of various theoretical papers dealing with adaptive minimax bounds in nonparametric estimation (see for instance Goldenshluger and Nemirovski [1997], Mathé [2006], Chichignoud [2012]). From the practical point of view, Lepski's method has also received further development, such as the intersection of confidence intervals (ICI) rule (see Katkovnik [1999]). This algorithm reveals computational advantages in comparison to the traditional Lepski's procedure, or even traditional cross-validation techniques since it does not require to compute all the estimators of the family. It was originally designed for a problem of gaussian filtering, which is at the core of many applications in image processing (see Kervrann and Boulanger [2006], Astola, Egiazarian, Foi, and Katkovnik [2010] and references therein). In a deconvolution setting as well, Comte and Lacour [2013] obtain adaptive optimal results (for pointwise and global risks) using an improvement of the standard Lepski's principle (see also Goldenshluger and Lepski [2011]).

In this chapter, we investigate the problem of bandwidth selection in empirical risk minimizations. Section 3.1 could be considered as a first step into the study of data-driven selection rule for the problem of inverse statistical learning. By considering empirical risks instead of estimators, Lepski's heuristic allows to select an isotropic bandwidth in noisy clustering thanks to ERC (Empirical Risk Comparison) method. In Section 3.2, we want to deal with a more challenging problem : the general bandwidth selection in kernel empirical risk minimization with anisotropic regularity assumptions. These could be done by extending the Goldenshluger-Lepski procedure in the same way as we extend the Lepski's method. The proposed method is called EGC (Empirical Gradient Comparison). However, as we will see, due to the presence of fast rates and a localization technique, this problem needs the introduction of a new criterion : the gradient excess risk. It allows to obtain adaptive minimax rates of convergence in a variety of nonparametric models where a bandwidth needs to be selected in an empirical risk minimization problem. Eventually, in Section 3.3, we compute the two proposed methods ERC and EGC in noisy clustering. It illustrates rather well the theoretical results of Section 3.1-3.2 below.

3.1 Isotropic case : the ERC method [L6]

This section is devoted to the problem of adaptive noisy clustering. We design a new selection rule based on the Lepski's principle with a comparison of *empirical risks* with different nuisance parameters. This method, called *Empirical Risk Comparison* (ERC), allows us to derive adaptive fast rates of convergence.

To the best of our knowledge, standard adaptive procedures such as cross-validation, model selection or aggregation cannot be directly applied in this particular context. In supervised learning (such as regression or binary classification), it is standard to choose a bandwidth - or a tuning - parameter using a decomposition of the set of observations. A training set is used to construct a family of candidate estimators, each one associated with a different value of the bandwidth. Then, a test set allows to estimate the generalization performances of each candidate. It gives rise to the family of cross-validation methods, or aggregation procedures. Unfortunately, in unsupervised tasks, this simple estimation is not possible. The lack of efficiency of cross-validation methods in clustering has been illustrated in Hastie, Tibshirani, and Friedman [2002] for the problem of choosing k in the k-means. In the presence of errors in variables, such as in deconvolution, it is quite obvious to perform cross-validation to choose the bandwidth of a deconvolution estimator. As described in Meister [2009], it is possible to estimate the squared risk $|\hat{f}_h - f|^2$ with Plancherel theorem, leading to the estimation of the Fourier transform of the unknown density. However, in our framework, this method seems hopeless since the optimal value of h does not minimize a squared risk but an excess risk. Eventually, model selection was introduced for selecting the hypothesis space over a sequence of nested models (e.g. finite dimension models) with a fixed empirical risk. Penalization methods are also suitable to choose smoothing parameters of well-known statistical methods such as splines, SVM or Tikhonov regularization methods. The idea is to replace the choice of the smoothing parameter by the choice of the radius into a suitable ellipsoid. Unfortunately, here, the nuisance parameter h affects directly the empirical risk and a model selection method can not be applied in this context.

We first recall the problem of clustering noisy data introduced in Section 2.2.3. Suppose we observe a corrupted sample Z_i, $i = 1, \dots, n$ of i.i.d. observations satisfying :

$$Z_i = X_i + \epsilon_i, \ i = 1, \dots, n. \tag{3.3}$$

We denote by f the unknown density (with respect to the Lebesgue measure on $\mathbb{R}^d$) of the i.i.d. sequence $X_1, X_2, ..., X_n$ and η the known density of the i.i.d. random variables $\epsilon_1, \epsilon_2, ..., \epsilon_n$, independent of the sequence $(X_i)_{i=1}^n$. We also assume that $X_1 \in \mathcal{B}(0,1)$ almost surely, where $\mathcal{B}(0,1)$ is the unit Euclidean ball of $\mathbb{R}^d$ (extension to $\mathcal{B}(0,M)$ with $M > 1$ is straightforward). Given some integer $k \geq 1$, the problem of noisy clustering consists in learning k clusters from f when a contaminated empirical version $Z_1, \dots, Z_n$ is observed. This problem is a particular case of inverse statistical learning which has deserved particular attention in Chapter 2, where non-adaptive results are proposed for a collection of noisy k-means minimizers :

$$\hat{\mathbf{c}}_h := \arg\min_{\mathbf{c} \in \mathcal{C}} \widehat{R}_h(\mathbf{c}), \quad h > 0, \tag{3.4}$$

where $\mathcal{C} := \{\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk} : c_j \in \mathcal{B}(0,1),\ j = 1, \ldots, k\}$ is the set of possible codebooks and $\widehat{R}_h(\mathbf{c})$ is defined according to :

$$\widehat{R}_h(\mathbf{c}) = \int_{\mathcal{B}(0,1)} \ell(\mathbf{c}, x)\hat{f}_h(x)dx = \frac{1}{n}\sum_{i=1}^{n} \ell_h(\mathbf{c}, Z_i), \tag{3.5}$$

where $\hat{f}_h(\cdot)$ is the deconvolution kernel estimator introduced in Chapter 2. In (3.5), $\ell_h(\mathbf{c}, Z)$ is the following convolution product :

$$\ell_h(\mathbf{c}, Z) := \left[\widetilde{\mathcal{K}}_h * (\ell(\mathbf{c}, \cdot)\mathbf{1}_{\mathcal{B}(0,1)}(\cdot))\right](Z) = \int_{\mathcal{B}(0,1)} \widetilde{\mathcal{K}}_h(Z - x)\ell(\mathbf{c}, x)dx, \quad \mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{C},$$

where $\ell(\mathbf{c}, x)$ is the standard k-means loss function.

The parameter h in (3.4)-(3.5) is of great interest in this chapter. In particular, an appropriate choice of the bandwidth provides in Chapter 2 fast rates thanks to the following bias-variance decomposition :

$$\begin{aligned} R(\hat{\mathbf{c}}_h, \mathbf{c}^\star) \le (R - \widehat{R}_h)(\hat{\mathbf{c}}_h, \mathbf{c}^\star) &\le (R - R_h)(\hat{\mathbf{c}}_h, \mathbf{c}^\star) + (R_h - \widehat{R}_h)(\hat{\mathbf{c}}_h, \mathbf{c}^\star) \\ &=: \text{bias}(h) + \text{var}(h), \end{aligned} \tag{3.6}$$

where in the sequel, for any fixed $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$, $R_h(\mathbf{c}, \mathbf{c}') := \mathbb{E}\left[\widehat{R}_h(\mathbf{c}) - \widehat{R}_h(\mathbf{c}')\right]$, $\mathbb{E}$ is the expectation w.r.t. the law of $(Z_1, \ldots, Z_n)$ and $\mathbf{c}^\star \in \mathcal{M}$, where $\mathcal{M}$ is the (finite) set of minimizers of $R(\mathbf{c})$. The first part of the decomposition is called a *bias* term, which depends on the unknown smoothness $\gamma > 0$ of the density f and on the deconvolution kernel. The second term of this decomposition is called the *variance* term, which is the stochastic error of the empirical risk minimization. It depends on a complexity parameter and on the noise assumption. It was controlled in Theorem 7 using empirical process theory in the spirit of Blanchard, Bousquet, and Massart [2008]. As a first step, we derive in Section 3.1.1 optimal fast rates of convergence with the bandwidth $\bar{h} := \bar{h}(\gamma)$ which minimizes the latter bias-variance trade-off (see Corollary 2).

3.1.1 Non-adaptive risk bound

In this paragraph, we write a non-adaptive excess risk bound for the noisy k-means procedure (3.4) in the isotropic framework. It is a particular case of Theorem 7 when the density f has a similar behaviour in each direction.

Corollary 2. *Assume that* $\mathbf{NA}(\rho, \beta)$ *and* ***PRC*** *of Chapter 2 are satisfied for some* $\beta \in (1/2, \infty)^d$ *and* $\rho > 0$. *Suppose* $\eta_\infty := \|\eta\|_\infty < \infty$ *and* $f \in \Sigma(\gamma, L)$ *with* $\gamma, L > 0$. *Then, if we consider* $\hat{\mathbf{c}}_{\bar{h}}$ *defined in* (3.4) *with :*

$$\bar{h} = n^{-1/(2\gamma + 2\bar{\beta})},$$

there exists a universal constant C_1 *depending on* $w, L, d, \gamma, \beta, \rho, k, \eta_\infty$ *and* $|\mathcal{M}|$, *and an integer* $n_0 \in \mathbb{N}^\star$ *such that for any* $\mathbf{c}^\star \in \mathcal{M}$ *and any* $n \ge n_0$ *:*

$$\mathbb{E}R(\hat{\mathbf{c}}_{\bar{h}}, \mathbf{c}^\star) \le C_1 n^{-\gamma/(\gamma + \bar{\beta})},$$

where $\bar{\beta} = \sum_{v=1}^{d} \beta_v$.

The proof is similar to the proof of Theorem 7 in Chapter 2 applied to the isotropic case.

3.1.2 Bandwidth selection with ERC

We turn out into the data-driven choice of the bandwidth $h > 0$ in the collection of estimators $\{\hat{\mathbf{c}}_h, h > 0\}$ defined in (3.4). The goal is to reach adaptive excess risk bound similar to Corollary 2 for a choice of h which does not depend on the smoothness of f.

Corollary 2 above motivates the use of a comparison method based on Lepski's principle (Lepski [1990]). Indeed, the non-adaptive choice of $\bar{h} = n^{-1/(2\gamma + 2\bar{\beta})}$ trades off a bias-variance decomposition of the excess risk (see (3.6)) and allows to get fast rates of convergence. As a result, Lepski's principle appears as the most suitable tool to construct an adaptive estimator $\hat{\mathbf{c}}_{\hat{h}}$, where $\hat{h}$ mimics the oracle $\bar{h}$ of Corollary 2. The construction of the data-driven bandwidth is based on the comparison of *empirical risks* instead of estimators. This direction has been already investigated in Polzehl and Spokoiny [2006], in a particular case of Kullback-Leibler divergence. In the sequel, we adopt the same point of view in noisy clustering by comparing empirical risks (3.5) with different bandwidths. The built estimator $\hat{\mathbf{c}}_{\hat{h}}$ will be called adaptive since it does not depend on the smoothness γ.

To define the selection rule, we first remind some definitions and notations. Given a kernel $\mathcal{K}$ satisfying the previous *Kernel assumption* (see Chapter 2), we note $\|\mathcal{K}\|_1$ the L_1-norm of the kernel on $\mathbb{R}^d$. The constant $\eta_\infty := \|\eta\|_\infty$ is the sup-norm of the noise density η, whereas $\rho > 0$ and $\bar{\beta} = \sum_{v=1}^d \beta_v$ are parameters involved in the noise assumption $\mathbf{NA}(\rho, \beta)$. Moreover, $C_2 > 0$ is the constant introduced in Lemma 4 of Chapter 2. It could be related to a familiar margin assumption. In the sequel, $\mathcal{V}(d) = \pi^{d/2}/\Gamma(d/2+1)$, where $\Gamma(\cdot)$ stands for the Gamma function.
Define the threshold term :

$$\delta_h := \frac{2^{10}\sqrt{2}\mathcal{V}(d)\|\mathcal{K}\|_1^2 C_2 \eta_\infty}{\rho^2} \frac{h^{-2\bar{\beta}}\log(n)}{n}, \tag{3.7}$$

where h belongs to the bandwidth set $\mathcal{H} := [h_{\min}, h_{\max}]$ with

$$h_{\min} := \frac{\log^{1/\bar{\beta}}(n)}{n^{1/2\bar{\beta}}} \text{ and } h_{\max} := \big(1/\log(n)\big)^{1/(2\gamma^+ + 2\bar{\beta})},$$

where $\gamma^+ > 0$ is an upper bound on the regularity index of f. In this section, we take n sufficiently large such that $n^{-1/(2\gamma+2\bar{\beta})} \in \mathcal{H}$. Moreover, for some constant $a \in (0,1)$, we set :

$$h_a := \{h \in \mathcal{H} : \exists m \in \mathbb{N}\,,\; h = h_{\max} a^m\}\,,$$

a discrete exponential net on the bandwidth set with cardinality $|h_a|$.

We are ready to introduce the adaptive bandwidth choice, called *Empirical Risk Comparison* (ERC) :

$$\hat{h} = \max\Big\{h \in h_a \;:\; \widehat{R}_{h'}(\hat{\mathbf{c}}_h) - \widehat{R}_{h'}(\hat{\mathbf{c}}_{h'}) \le 3\delta_{h'},\; \forall h' \le h\Big\}. \tag{3.8}$$

The noisy k-means estimator (2.37) with bandwidth $\hat{h}$ chosen from ERC rule (3.8) has the following property.

Theorem 9. *Assume that* $\mathbf{NA}(\rho, \beta)$ *and* **PRC** *are satisfied for some* $\beta \in (1/2, \infty)^d$, $\rho > 0$. *Suppose* $\eta_\infty := n^o \eta n^o{}_\infty < \infty$ *and* $f \in \Sigma(\gamma, L)$, *where* $\gamma \in [0, \gamma^+)$ *and* $L > 0$. *Then, there exists a universal constant depending on* $C_2, w, L, d, \gamma, \beta, \rho, k, \eta_\infty$, $|\mathcal{M}|$, *and* $n_1 \in \mathbb{N}$ *such that for any* $\mathbf{c}^\star \in \mathcal{M}$ *and any* $n \ge n_1$, *estimator* $\hat{\mathbf{c}}_{\hat{h}}$ *with* $\hat{h}$ *selected by ERC rule* (3.8) *satisfies :*

$$\mathbb{E}R(\hat{\mathbf{c}}_{\hat{h}}, \mathbf{c}^\star) \le C_3 \left(\frac{\log(n)}{n}\right)^{\gamma/(\gamma+\bar{\beta})},$$

where $\bar{\beta} = \sum_{v=1}^d \beta_v$ *and* $C_3 > 0$.

Theorem 9 is an adaptive upper bound for the estimator $\hat{\mathbf{c}}_{\hat{h}}$, where $\hat{h}$ is chosen from the ERC selection rule (3.8). The estimator $\hat{\mathbf{c}}_{\hat{h}}$ is then adaptive w.r.t. the smoothness γ. This adaptive excess risk bound coincides with the non-adaptive previous result of Corollary 2, up to an extra log term. This is the price to pay for the data-driven property of the procedure. A natural question is the optimality of Theorem 9 in the minimax sense.

In this respect, let us remind that it is standard from Lepski [1990] (see also Brown and Low [1996]) to pay a $\log(n)$ factor in pointwise estimation (i.e. when we estimate a target function f at a given point x_0). However, it is well-known that there is no price to pay for adaptivity in global estimation (e.g. in L_p-norm). In the problem of noisy clustering, or more generally noisy classification, the choice of h concerns the global estimation of the density f. This estimation is used in the procedure of noisy k-means, where we plug $\hat{f}_h$ into the true risk. We could conjecture that a global estimation of f is sufficient and thus no extra log term has to be paid. This is one of the challenge of the next section.

The threshold term δ_h - which comes from the control of the stochastic part of the excess risk - has the following form (see (3.7)) :

$$\delta_h = C_{\text{adapt}} \frac{h^{-2\bar{\beta}} \log(n)}{n},$$

where the (large) constant $C_{\text{adapt}} > 0$ depends on the model. Indeed, by definition, it depends on the underlying density f. In practice, we recommend a painstaking calibration of this constant. From the theoretical point of view, this constant could be chosen from the *propagation* method suggested by Spokoiny and Vial [2009].

The construction of an algorithm to compute the ERC rule is also of first interest. This is the core of Section 3.3, where an ICI rule is adapted for ERC. As in standard kernel estimation, the implementation of the ICI algorithm will be efficient to avoid the calculation of all the estimators in the collection of noisy k-means.

In this isotropic setting, we could propose extensions of the ERC selection rule (3.8) to a more general context of kernel empirical risk minimization. However, we prefer to defer these considerations to Section 3.2, where we move to the anisotropic case in a more general context of kernel empirical risk minimization.

3.2 Anisotropic case : the gradient [L7]

As seen before, Lepski-type procedures are rather appropriate to construct data-driven bandwidths involved in kernels. Nevertheless, it is well-known that these procedures suffer from the restriction to isotropic bandwidths with multidimensional data, which is the consideration of nested neighborhoods (hyper-cube). Many improvements have been made by Kerkyacharian et al. [96] and more recently by Goldenshluger and Lepski [73] to select anisotropic bandwidths (hyper-rectangle). However, theses approaches still do not provide anisotropic bandwidth selection for non-linear estimators as in our purpose. The only work we can mention is Chichignoud and Lederer [48] in a restrictive case which is pointwise estimation in nonparametric regression. Therefore, the study of data-driven selection of anisotropic bandwidths deserves some clarifications. Moreover, this field is of first interest in practice, especially in image denoising (see e.g. Arias-Castro, Salmon, and Willett [2012], Astola, Egiazarian, Foi, and Katkovnik [2010]).

This section tries to fill this gap in the context of *kernel empirical risk minimization*. We consider the minimization problem of an unknown risk function $R : \mathbb{R}^m \to \mathbb{R}$, where $m \geq 1$ is the dimension of the statistical model we have at hand[1]. Assume there exists a risk minimizer :

$$\theta^\star \in \arg\min_{\theta \in \mathbb{R}^m} R(\theta). \tag{3.9}$$

The risk function corresponds to the expectation of an appropriate loss function w.r.t. an unknown distribution. In empirical risk minimization, this quantity is usually estimated by its empirical version from an i.i.d. sample. However, in many problems such as local M-estimation or errors-in-variables models, a nuisance parameter can be involved in the empirical version. This parameter most often coincides with some bandwidth related to a kernel which gives rise to the problem of kernel empirical risk minimization. One typically deals with this issue in pointwise estimation as e.g. in Polzehl and Spokoiny [142] with localized likelihoods or in Chichignoud and Lederer [48] in the setting of robust estimation with local M-estimators. In this manuscript, we have investigated supervised and unsupervised learning with errors in variables. As a rule, such issues (viewed as an inverse problem) require to plug-in deconvolution

[1] In (3.9), we consider the risk minimization over a finite dimensional space $\mathbb{R}^m$. In statistical learning or nonparametric estimation, one usually aims at estimating a functional object belonging to some Hilbert space. However, in many examples, the target function can be approximated by a finite object thanks to a suitable decomposition in a basis of the Hilbert space for instance. This is typically the case in local M-estimation, where the target function is assumed to be locally polynomial (and even constant in many cases). Moreover, in statistical learning, one is often interested in the estimation of a finite number of parameters as in clustering. The extension to the infinite dimensional case is discussed at the end of the manuscript.

kernels in the empirical risk (see Chapter 2 for various examples). The choice of the bandwidth is therefore one of the biggest challenges. In this respect, data-driven bandwidth selection in the isotropic case has been considered in Section 3.1.

In this section, we provide a novel universal data-driven selection of anisotropic bandwidths suitable for our large context of models. This method can be viewed as a generalization of the so-called Goldenshluger-Lepski method (GL method, see Goldenshluger and Lepski [2011]) and of the Empirical Risk Comparison method of the previous section. We especially derive an oracle inequality for the "Gradient excess risk" (described below), which leads to adaptive optimal results in many settings such as pointwise and global estimation in nonparametric regression and clustering with errors-in-variables.

3.2.1 The gradient excess risk approach

Along the present section, we deal with smooth loss functions, where the smoothness is related to the differentiability of the associated risk function. Under this restriction, we propose a new criterion to measure the performance of an estimator $\widehat{\theta}$, namely the *Gradient excess risk* (G-excess risk for short in the sequel). This quantity is defined as :

$$|G(\widehat{\theta},\theta^\star)|_2 := |G(\widehat{\theta}) - G(\theta^\star)|_2 \text{ where } G := \nabla R, \tag{3.10}$$

where $|\cdot|_2$ denotes the Euclidean norm in $\mathbb{R}^m$ and $\nabla R : \mathbb{R}^m \to \mathbb{R}^m$ denotes the gradient of the risk R. With a slight abuse of notation, G denotes the gradient, whereas $G(\cdot,\theta^\star)$ denotes the G-excess risk. The use of a smooth loss function, together with (3.9), leads to $G(\theta^\star) = (0,\dots,0)^\top \in \mathbb{R}^m$ and the G-excess risk $|G(\theta,\theta^\star)|_2$ corresponds to $|G(\theta)|_2$. The most important fact with (3.10) is the following one : with smooth loss functions, slow rates $\mathcal{O}(n^{-1/2})$ for the G-excess risk $|G(\widehat{\theta},\theta^\star)|_2$ lead to fast rates $\mathcal{O}(n^{-1})$ for the usual excess risk $R(\widehat{\theta}) - R(\theta^\star)$ thanks to the following lemma.

Lemma 5. *Let $\theta^\star$ satisfy* (3.9) *and U be the Euclidean ball of $\mathbb{R}^m$ centered at $\theta^\star$, with radius $\delta > 0$. Assume $\theta \mapsto R(\theta)$ is $\mathcal{C}^2(U)$, all of second partial derivatives of R are bounded on U by a constant κ_1 and the Hessian matrix $H_R(\cdot)$ is positive definite at $\theta^\star$. Then, for $\delta > 0$ small enough, we have :*

$$\sqrt{R(\theta) - R(\theta^\star)} \leq 2\frac{\sqrt{m\kappa_1}}{\lambda_{\min}}|G(\theta,\theta^\star)|_2, \ \forall \theta \in U,$$

where $\lambda_{\min}$ is the smallest eigenvalue of $H_R(\theta^\star)$.

The proof is based on the inverse function theorem and a simple Taylor expansion of the function $R(\cdot)$. The constant two appearing in the RHS can be arbitrarily close to one, depending on the size of the neighborhood.

Let us explain how the previous lemma, together with standard probabilistic tools, allows us to establish fast rates for the excess risk. Recall $\widehat{R}$ denotes the usual empirical risk with associated gradient $\widehat{G} := \nabla \widehat{R}$ and associated empirical risk minimizer (ERM) $\widehat{\theta}$ for ease of exposition. Under a smoothness hypothesis over the loss function, $G(\theta^\star) = \widehat{G}(\widehat{\theta}) = (0,\dots,0)^\top$ and we lead to the following heuristic [2] :

$$\sqrt{R(\widehat{\theta}) - R(\theta^\star)} \lesssim |G(\widehat{\theta},\theta^\star)|_2 = |G(\widehat{\theta}) - \widehat{G}(\widehat{\theta})|_2 \leq \sup_{\theta \in \mathbb{R}^m} |G(\theta) - \widehat{G}(\theta)|_2 \lesssim n^{-1/2}. \tag{3.11}$$

The last inequality comes from the application of a concentration inequality to the empirical process $\widehat{G}(\cdot)$, which requires no localization technique. Somehow, Lemma 5 guarantees that for a smooth loss function, fast rates occur when the Hessian matrix of the risk is positive definite at $\theta^\star$.

Now, let us compare our approach to the literature on excess risk bounds. Vapnik and Chervonenkis [174] have originally proposed to control the excess risk via the theory of empirical processes. It gives rise to slow rates $\mathcal{O}(n^{-1/2})$ for the excess risk (see also Vapnik [1998]). In the last decade, many authors have improved such a bound by giving fast rates $\mathcal{O}(n^{-1})$ using the so-called localization technique (see Mammen and Tsybakov [1999],Koltchinskii and Panchenko [2000],Tsybakov [2004],Blanchard, Bousquet, and Massart [2008], Blanchard, Lugosi, and Vayatis [2003], Koltchinskii [2006] , Massart and Nédélec [2006],

[2] This idea (and precisely the equality in the middle of (3.11)) was initiated in Huber [1964] for robust estimation.

Mendelson [2003], and the references therein). This field has been especially studied in classification (see Boucheron et al. [27] for a nice survey). This complicated modus operandi requires a variance-risk correspondence, equivalent to the so-called margin assumption. Interestingly enough, the next lemma suggests to link the margin assumption with some smoothness conditions on the loss function as follows.

Lemma 6. *Let $\mathcal{X}$ be a $\mathbb{R}^p$-random variable with law $P_{\mathcal{X}}$ and assume there exists a loss function $\ell : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}_+$ such that $R(\cdot) = \mathbb{E}_{P_{\mathcal{X}}}\ell(\mathcal{X}, \cdot)$. Let us consider an oracle defined in* (3.9) *and let U be the Euclidean ball of center $\theta^\star$ and radius $\delta > 0$ such that :*

- *$\theta \mapsto \ell(\mathcal{X}, \theta)$ is twice differentiable on U, $P_{\mathcal{X}}$-almost surely;*
- *$R(\cdot) = \mathbb{E}\ell(\mathcal{X}, \cdot)$ is three times differentiable on U and the partial derivatives of third order are bounded;*
- *the Hessian matrix $H_R(\theta^\star)$ is positive definite.*

Then, for δ sufficiently small, we have :

$$\mathbb{E}_{P_{\mathcal{X}}} \left[\ell(\mathcal{X}, \theta) - \ell(\mathcal{X}, \theta^\star)\right]^2 \leq 3\kappa_1 \lambda_{\min}^{-1} \left[R(\theta) - R(\theta^\star)\right], \ \forall \theta \in U,$$

where $\kappa_1 = \mathbb{E}_{P_{\mathcal{X}}} \sup_{\theta \in U} |\nabla \ell(\mathcal{X}, \theta)|_2^2$ and $\lambda_{\min}$ is the smallest eigenvalue of $H_R(\theta^\star)$.

Note that the regularity of the loss function implies a strong margin assumption, i.e. a power of the excess risk equals to 1 in the RHS. Weaker margin assumptions - where the power of the excess risk is less than 1 - have been considered in the literature (see Tsybakov [165], Koltchinskii [2006], Bartlett and Mendelson [2006]) and allow them to obtain fast rates of convergence for the excess risk between $\mathcal{O}(n^{-1/2})$ and $\mathcal{O}(n^{-1})$. However, to the best of our knowledge, these weaker margin assumptions are very often related to non-smooth loss functions, such as the hinge loss or the hard loss in the specific context of binary classification.

From the model selection point of view, standard penalization techniques - based on localization - suffer from the dependency on parameters involved in the margin assumption. More precisely, in the strong margin assumption framework, the construction of the penalty needs the knowledge of $\lambda_{\min}$, related to the Hessian matrix of the risk. This constant especially coincides with the usual Fisher information in maximum likelihood estimation. Although many authors have recently investigated the adaptivity w.r.t. these parameters, by proposing "margin-adaptive" procedures (see Polzehl and Spokoiny [2006] for the propagation method, Lecué [2007] for aggregation and Arlot and Massart [2009] for the slope heuristic), the theory is not completed and remains a hard issue (see the related discussion in Section 3.2.7). As an alternative, it is surprising to note that our data-driven procedure does not suffer from the dependency on $\lambda_{\min}$ since we focus on the G-excess risk.

3.2.2 Application : fast rates in clustering

As an illustration, we expound fast rates of convergence in clustering based on the gradient excess risk approach. For this purpose, we go back to the classical statistical learning problem of clustering. Let us consider an integer $k \geq 1$ and a $\mathbb{R}^d$-random variable X with law P with density f w.r.t. the Lebesgue measure on $\mathbb{R}^d$ satisfying $\mathbb{E}_P|X|_2^2 < \infty$, where $|\cdot|_2$ stands for the Euclidean norm in $\mathbb{R}^d$. We restrict the study to $[0,1]^d$, assuming that $X \in [0,1]^d$ almost surely. In the sequel, $\mathbf{c} = (c_1, \ldots, c_k) \in (\mathbb{R}^d)^k$ is a set of codebooks. Then, we want to construct a codebook $\mathbf{c}$ minimizing some risk or distortion :

$$R(\mathbf{c}) := \mathbb{E}_P \ell(\mathbf{c}, X), \tag{3.12}$$

where $\ell(\mathbf{c}, x)$ measures the loss of the codebook $\mathbf{c}$ at point x. For ease of exposition, we study the risk minimization of (3.12) based on the Euclidean distance, by choosing a loss function related to the standard k-means loss function, namely :

$$\ell(\mathbf{c}, x) = \min_{j=1,\ldots,k} |x - c_j|_2^2, \quad x \in \mathbb{R}^d.$$

In the direct case, we have at our disposal an i.i.d. sample $(X_1, \ldots, X_n)$ with law P and an associated ERM :

$$\widehat{\mathbf{c}} \in \arg\min_{\mathbf{c} \in \mathbb{R}^{dk}} \widehat{R}(\mathbf{c}), \text{ where } \widehat{R}(\mathbf{c}) := \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{c}, X_i). \tag{3.13}$$

Recently, Levrard [117] has proved fast rates of convergence $\mathcal{O}(n^{-1})$ under a margin assumption. The first message of this paragraph is to highlight that using the gradient of (3.12), similar results could be proved with a significantly simpler proof.

For this purpose, we assume that the Hessian matrix H_R is positive definite at each oracle $\mathbf{c}^\star$. As viewed in the previous chapter, this assumption has been considered for the first time in Pollard [139] and is often referred as the Pollard's regularity assumptions. Under this assumption, we can state the same kind of result as Lemma 5 in the framework of clustering with k-means.

Lemma 7. *Let $\boldsymbol{c}^\star$ be a minimizer of* (3.12) *and assume $H_R(\boldsymbol{c}^\star)$ is positive definite. Let us consider $\mathbb{C} := \{\boldsymbol{c} = (c_1, \dots, c_k) \in [0,1]^{dk} : \forall i \neq j \in \{1, \dots, k\}, c_i \neq c_j\}$. Then :*

- *$\forall x \in \mathbb{R}^d$, $\boldsymbol{c} \mapsto \ell(\boldsymbol{c}, x)$ is infinitely differentiable on $\mathbb{C} \setminus \Delta_x$, where $\Delta_x = \{c \in [0,1]^{dk} : x \in \partial V(\boldsymbol{c})\}$ and $\partial V(\boldsymbol{c}) = \{x \in \mathbb{R}^d : \exists i \neq j$ such that $|x - c_i|_2 = |x - c_j|_2\}$;*
- *Let U be the Euclidean ball center at $\boldsymbol{c}^\star$ with radius $\delta > 0$. Then, for δ sufficiently small :*

$$\sqrt{R(\boldsymbol{c}) - R(\boldsymbol{c}^\star)} \leq 2\frac{\sqrt{2kd}}{\lambda_{\min}} |G(\boldsymbol{c}, \boldsymbol{c}^\star)|_2, \ \forall \boldsymbol{c} \in U,$$

where $\lambda_{\min} > 0$ is the smallest eigenvalue of $H_R(\boldsymbol{c}^\star)$.

As mentioned above, we need the consistency - in terms of Euclidean distance - of the ERM $\widehat{\mathbf{c}}$ defined in (3.13) in order to obtain the inequality of Lemma 7 with $\mathbf{c} = \widehat{\mathbf{c}}$. Pollard [139] has especially studied the consistency of $\widehat{\mathbf{c}}$ and allows us to satisfy our needs :

Theorem 10. *Suppose the assumptions of Lemma 7 hold. Then, for n sufficiently large, the ERM $\widehat{\boldsymbol{c}}$ defined in* (3.13) *satisfies :*

$$\mathbb{E}R(\widehat{\boldsymbol{c}}) - R(\boldsymbol{c}^\star) \leq \frac{8b_1^2 kd\lambda_{\min}^{-2}}{n},$$

where $b_1 > 0$ is a constant.

The proof is a direct application of the heuristic (3.11). In particular, the study of an empirical process based on the gradient leads to slow rates $\mathcal{O}(n^{-1/2})$ for the G-excess risk. Lemma 7 concludes the proof.

3.2.3 Kernel empirical risk minimization and examples

In this section, we are primarily interested in the kernel empirical risk minimization problem, where a bandwidth is involved in the empirical risk. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and for some $p \in \mathbb{N}^*$, consider a $\mathbb{R}^p$-random variable Z on $(\Omega, \mathcal{F}, \mathbb{P})$ with law P absolutely continuous w.r.t. the Lebesgue measure. In what follows, we observe a sample $\mathcal{Z}_n := \{Z_1, \dots, Z_n\}$ of independent and identically distributed (i.i.d.) random variables with law P. The expectation w.r.t. the law of $\mathcal{Z}_n$ is denoted by $\mathbb{E}$. Moreover, in the sequel, for some $d \in \mathbb{N}^*$, we consider a kernel $\mathcal{K}_h : \mathbb{R}^d \to \mathbb{R}$ of order $r \in \mathbb{N}^d$ and define the kernel empirical risk indexed by an anisotropic bandwidth $h \in \mathcal{H} \subset (0,1]^d$ as :

$$\widehat{R}_h(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathcal{K}_h}(Z_i, \theta), \tag{3.14}$$

and an associated kernel empirical risk minimizer (kernel ERM) :

$$\widehat{\theta}_h \in \arg\min_{\theta \in \mathbb{R}^m} \widehat{R}_h(\theta). \tag{3.15}$$

In the sequel, the function $\ell_{\mathcal{K}_h} : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}_+$ is a loss function associated to a kernel $\mathcal{K}_h$ such that $\theta \mapsto \ell_{\mathcal{K}_h}(Z, \theta)$ is twice differentiable P almost surely and such that $\widehat{R}_h$ is an asymptotically unbiased estimator of the true risk R, i.e.

$$\lim_{h \to (0,\dots,0)} \mathbb{E}\widehat{R}_h(\theta) = \mathbb{R}(\theta), \ \forall \theta \in \mathbb{R}^m. \tag{3.16}$$

The agenda is the data-driven selection of the "best" kernel ERM in the family $\{\widehat{\theta}_h, h \in \mathcal{H}\}$. This problem arises in many examples, such as local fitted likelihood (Polzehl and Spokoiny [2006]), image denoising

(Astola, Egiazarian, Foi, and Katkovnik [2010]), or robust nonparametric regression (Chichignoud and Lederer [2013]). In such a framework, we observe a sample of i.i.d. pairs $Z_i = (W_i, Y_i)_{i=1}^n$ and the kernel empirical risk has the following general form :

$$\frac{1}{n}\sum_{i=1}^n \ell_{\mathcal{K}_h}(Z_i,\theta) = \frac{1}{n}\sum_{i=1}^n \rho(Z_i,\theta)\mathcal{K}_h(W_i - x_0),$$

where $\rho(\cdot,\theta)$ is some likelihood whereas $\mathcal{K}_h(\cdot)$ is a standard kernel function. Of course, we can also go back to the previous inverse statistical learning context, where a deconvolution kernel $\widetilde{\mathcal{K}}_h(\cdot)$ is involved in the empirical risk, such as in (3.5).

In the sequel, we present the selection rule in the general context of kernel empirical risk minimization. We especially deal with the noisy clustering in Section 3.2.5 whereas Section 3.2.6 is dedicated to robust nonparametric regression.

3.2.4 General oracle inequality

The anisotropic bandwidth selection problem has been recently investigated in Goldenshluger and Lepski [73] (GL method) in density estimation (see also Comte and Lacour [2013] in deconvolution estimation and Goldenshluger and Lepski [2008], Goldenshluger and Lepski [2009] for the white noise model). This method, based on the comparison of estimators, requires some "linearity" property, which is trivially satisfied for kernel estimators in density estimation. However, kernel ERM are usually non-linear (except for the least square estimator), and the GL method cannot be directly applied. A first trail would be to compare the empirical risks (3.14) - viewed as estimators - instead of minimizers. This comparison has been already employed with the ERC method, which is only suitable for isotropic bandwidths. Unfortunately, as far as we know, the GL method cannot be performed by using this comparison. More precisely, the requirement of the localization argument seems to be the main obstacle to the GL method.

To tackle this impasse, we introduce a new selection rule based on the comparison of gradient empirical risks instead of kernel ERM (i.e. estimators as in Goldenshluger and Lepski [2011]). For any $h \in \mathcal{H}$ and any $\theta \in \mathbb{R}^m$, the gradient empirical risk is defined as :

$$\widehat{G}_h(\theta) := \frac{1}{n}\sum_{i=1}^n \nabla\ell_{\mathcal{K}_h}(Z_i,\theta) = \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta_j}\ell_{\mathcal{K}_h}(Z_i,\theta)\right)_{j=1,\dots,m}. \tag{3.17}$$

Note that we have coarsely $\widehat{G}_h(\widehat{\theta}_h) = (0,\dots,0)^\top$ since $\ell_{\mathcal{K}_h}(Z_i,\cdot)$ is twice differentiable almost surely. According to (3.16), we also notice that the G-empirical risk is an asymptotically unbiased estimator of the gradient of the risk.

As mentioned in Chapter 1, we need to introduce an auxiliary G-empirical risk in the comparison. For any couple of bandwidths $(h,h') \in \mathcal{H}^2$ and any $\theta \in \mathbb{R}^m$, the auxiliary G-empirical risk is defined as :

$$\widehat{G}_{h,h'}(\theta) := \frac{1}{n}\sum_{i=1}^n \nabla\ell_{\mathcal{K}_h*\mathcal{K}_{h'}}(Z_i,\theta), \tag{3.18}$$

where $\mathcal{K}_h*\mathcal{K}_{h'}(\cdot) := \int_{\mathbb{R}^d}\mathcal{K}_h(\cdot - x)\mathcal{K}_{h'}(x)dx$ stands for the convolution between $\mathcal{K}_h$ and $\mathcal{K}_{h'}$. The statement of the main oracle inequality needs a control of the deviation of some random processes depending on the auxiliary G-empirical risk. This control is given by the next definition.

Definition 7 (Majorant)**.** *For any integer $l > 0$, we call majorant a function $\mathcal{M}_l : \mathcal{H}^2 \to \mathbb{R}_+$ such that :*

$$\mathbb{P}\left(\sup_{h,h'\in\mathcal{H}}\left\{|\widehat{G}_{h,h'} - \mathbb{E}\widehat{G}_{h,h'}|_{2,\infty} + |\widehat{G}_{h'} - \mathbb{E}\widehat{G}_{h'}|_{2,\infty} - \mathcal{M}_l(h,h')\right\}_+ > 0\right) \le n^{-l},$$

where $|T|_{2,\infty} := \sup_{\theta\in\mathbb{R}^m}|T(\theta)|_2$ for all $T : \mathbb{R}^m \to \mathbb{R}^m$ with $|\cdot|_2$ the Euclidean norm on $\mathbb{R}^m$.

The main issue for applications is to compute right order majorants. This could be done thanks to the empirical process theory, such as Talagrand's inequalities (see for instance Bousquet [2002], Goldenshluger and Lepski [2011]). In Section 3.2.5 and Section 3.2.6, such majorant functions are computed in noisy clustering and in robust nonparametric regression.

We are now ready to define the selection rule called *Empirical Gradient Comparison* (EGC) as :

$$\widehat{h} = \arg\min_{h\in\mathcal{H}} \widehat{\mathrm{BV}}(h), \tag{3.19}$$

where $\widehat{\mathrm{BV}}(h)$ is explicitly defined as :

$$\widehat{\mathrm{BV}}(h) := \sup_{h'\in\mathcal{H}} \left\{ |\widehat{G}_{h,h'} - \widehat{G}_{h'}|_{2,\infty} - \mathcal{M}_l(h,h') \right\} + \mathcal{M}_l^{\infty}(h), \quad \text{with } \mathcal{M}_l^{\infty}(h) := \sup_{h'\in\mathcal{H}} \mathcal{M}_l(h',h).$$

The function $\widehat{\mathrm{BV}}(\cdot)$ is based on the estimation of the following bias-variance decomposition :

$$|G(\widehat{\theta}_h,\theta^\star)|_2 \le \sup_{\theta\in\mathbb{R}^m} |G(\theta)-\widehat{G}_h(\theta)|_2 := |G-\widehat{G}_h|_{2,\infty} \le |\mathbb{E}\widehat{G}_h - G|_{2,\infty} + |\widehat{G}_h - \mathbb{E}\widehat{G}_h|_{2,\infty}, \tag{3.20}$$

where the expectation $\mathbb{E}$ is understood coordinatewise and $|T|_{2,\infty} := \sup_{\theta\in\mathbb{R}^m}|T(\theta)|_2$ for all functions $T:\mathbb{R}^m\to\mathbb{R}^m$. The selection rule is constructed in a way that the selected bandwidth mimics the oracle bandwidth $h^\star$, which trades off the bias-variance decomposition (3.20). The construction of $\widehat{\mathrm{BV}}(\cdot)$ consists of two steps : the variance (stochastic) term $|\widehat{G}_h - \mathbb{E}\widehat{G}_h|_{2,\infty}$ is controlled thanks to Definition 7 whereas the bias term $|\mathbb{E}\widehat{G}_h - G|_{2,\infty}$ is estimated with the auxiliary G-empirical risk (3.2.5).

The kernel ERM $\widehat{\theta}_{\widehat{h}}$ defined in (3.15) with bandwidth $\widehat{h}$ satisfies the following bound.

Theorem 11. *Let $\mathcal{M}_l(\cdot,\cdot)$ be a majorant according to Definition 7. For any $n\in\mathbb{N}^*$ and for any $l\in\mathbb{N}^*$, we have with probability $1-n^{-l}$:*

$$|G(\widehat{\theta}_{\widehat{h}},\theta^\star)|_2 \le 3\inf_{h\in\mathcal{H}}\left\{B(h)+\mathcal{M}_l^{\infty}(h)\right\},$$

where $B(\cdot):\mathcal{H}\to\mathbb{R}_+$ is a bias function defined as :

$$B(h) := \max\left(|\mathbb{E}\widehat{G}_h - G|_{2,\infty}, \sup_{h'\in\mathcal{H}} |\mathbb{E}\widehat{G}_{h,h'} - \mathbb{E}\widehat{G}_{h'}|_{2,\infty} \right), \quad \forall h\in\mathcal{H}.$$

Theorem 11 is the main result of this chapter. The G-excess risk of the data-driven estimator $\widehat{\theta}_{\widehat{h}}$ is bounded with high probability. Of course, a bound in expectation can be deduced coarsely. The proof of Theorem 11 (expounded below) is based on the definition of $\widehat{h}$ in (3.19). The first step is a decomposition of the G-excess risk by using the auxiliary G-empirical risk (3.18). Then, Definition 7 completes the proof.

The RHS in the oracle inequality can be viewed as the minimization of the usual bias-variance trade-off. Indeed, the bias term $B(h)$ is deterministic and tends to 0 as $h\to(0,\ldots,0)$. The sup-majorant $\mathcal{M}_l^{\infty}(h)$ upper bounds the stochastic part of the G-empirical risk and is viewed as a variance term.

Theorem 11 can be seen as an oracle inequality since minimizing the bias-variance trade-off in the RHS is sufficient to establish adaptive fast rates in noisy clustering and adaptive minimax rates in nonparametric estimation (see Sections 3.2.5 and 3.2.6).

In order to show the power of the G-excess risk, we simultaneously deduce a control of the estimation error $|\widehat{\theta}_{\widehat{h}} - \theta^\star|_2$ as well as a bound for the excess risk $R(\widehat{\theta}_{\widehat{h}}) - R(\theta^\star)$. In the presence of smooth loss functions, Lemma 5 is at the origin of the following corollary.

Corollary 3. *Suppose the assumptions of Lemma 5 are satisfied and for all $h\in\mathcal{H}$, the estimator $\widehat{\theta}_h$ of $\theta^\star$ is consistent. Then, for n sufficiently large, for any $l\in\mathbb{N}^*$, with probability $1-n^{-l}$, it holds :*

$$R(\widehat{\theta}_{\widehat{h}}) - R(\theta^\star) \le 36\frac{m\kappa_1}{\lambda_{\min}^2}\inf_{h\in\mathcal{H}}\left\{B(h)+\mathcal{M}_l^{\infty}(h)\right\}^2,$$

and

$$|\widehat{\theta}_{\widehat{h}} - \theta^\star|_2 \le 6\frac{\sqrt{m\kappa_1}}{\lambda_{\min}}\inf_{h\in\mathcal{H}}\left\{B(h)+\mathcal{M}_l^{\infty}(h)\right\},$$

where $\kappa_1,\lambda_{\min}$ are positive constants defined in Lemma 5.

We highlight that the consistency of all estimators $\{\widehat{\theta}_h,\ h \in \mathcal{H}\}$ is necessary in order to apply Lemma 5. This usually implies restrictions on the bandwidth set (see Sections 3.2.5 and 3.2.6 for further details).

The first inequality of Corollary 3 will be used in Section 3.2.5 in the setting of clustering with errors-in-variables. In this case, we are interested in excess risk bounds and the statement of fast rates of convergence. The second inequality of Corollary 3 is the main tool to establish minimax rates for both pointwise and global risks in the context of robust nonparametric regression (see Section 3.2.6).

The construction of the selection rule (3.19), as well as the upper bound in Theorem 11, does not suffer from the dependency of $\lambda_{\min}$ related to the smallest eigenvalue of the Hessian matrix of the risk (see Lemma 5). In other words, the method is robust w.r.t. this parameter, which is a major improvement in comparison with other adaptive or model selection methods of the literature.

Proof (of theorem 11): For some $h \in \mathcal{H}$, we start with the following decomposition :

$$\begin{aligned}|G(\widehat{\theta}_{\widehat{h}}, \theta^\star)|_2 &= \left|(\widehat{G}_{\widehat{h}} - G)(\widehat{\theta}_{\widehat{h}})\right|_2 \leq |\widehat{G}_{\widehat{h}} - G|_{2,\infty} \\ &\leq |\widehat{G}_{\widehat{h}} - \widehat{G}_{\widehat{h},h}|_{2,\infty} + |\widehat{G}_{\widehat{h},h} - \widehat{G}_h|_{2,\infty} + |\widehat{G}_h - G|_{2,\infty}.\end{aligned} \tag{3.21}$$

By definition of $\widehat{h}$ in (3.19), the first two terms in the RHS of (3.21) are bounded as follows :

$$\begin{aligned}|\widehat{G}_{\widehat{h}} - \widehat{G}_{\widehat{h},h}|_{2,\infty} + |\widehat{G}_{\widehat{h},h} - \widehat{G}_h|_{2,\infty} &= |\widehat{G}_{h,\widehat{h}} - \widehat{G}_{\widehat{h}}|_{2,\infty} - \mathcal{M}_\ell(h,\widehat{h}) + \mathcal{M}_\ell(\widehat{h},h) \\ &\quad + |\widehat{G}_{\widehat{h},h} - \widehat{G}_h|_{2,\infty} - \mathcal{M}_\ell(\widehat{h},h) + \mathcal{M}_\ell(h,\widehat{h}) \\ &\leq \sup_{h' \in \mathcal{H}} \left\{|\widehat{G}_{h,h'} - \widehat{G}_{h'}|_{2,\infty} - \mathcal{M}_\ell(h,h')\right\} + \mathcal{M}_\ell^\infty(h) \\ &\quad + \sup_{h' \in \mathcal{H}} \left\{|\widehat{G}_{\widehat{h},h'} - \widehat{G}_{h'}|_{2,\infty} - \mathcal{M}_\ell(\widehat{h},h')\right\} + \mathcal{M}_\ell^\infty(\widehat{h}) \\ &= \widehat{\mathrm{BV}}(h) + \widehat{\mathrm{BV}}(\widehat{h}) \leq 2\widehat{\mathrm{BV}}(h).\end{aligned} \tag{3.22}$$

Besides, the last term in (3.21) is controlled as follows :

$$\begin{aligned}|\widehat{G}_h - G|_{2,\infty} &\leq |\widehat{G}_h - \mathbb{E}\widehat{G}_h|_{2,\infty} + |\mathbb{E}\widehat{G}_h - G|_{2,\infty} \\ &\leq |\widehat{G}_h - \mathbb{E}\widehat{G}_h|_{2,\infty} - \mathcal{M}_l(h,h) + \mathcal{M}_l(h,h) + |\mathbb{E}\widehat{G}_h - G|_{2,\infty} \\ &\leq \sup_{h,h'} \left\{|\widehat{G}_{h,h'} - \mathbb{E}\widehat{G}_{h,h'}|_{2,\infty} + |\widehat{G}_{h'} - \mathbb{E}\widehat{G}_{h'}|_{2,\infty} - \mathcal{M}_l(h,h')\right\} \\ &\quad + \mathcal{M}_l^\infty(h) + |\mathbb{E}\widehat{G}_h - G|_{2,\infty} \\ &=: \zeta + \mathcal{M}_l^\infty(h) + |\mathbb{E}\widehat{G}_h - G|_{2,\infty}.\end{aligned}$$

Using (3.21) and (3.22), gathering with the last inequality, we have for all $h \in \mathcal{H}$:

$$|G(\widehat{\theta}_{\widehat{h}}, \theta^\star)|_2 \leq 2\widehat{\mathrm{BV}}(h) + \zeta + \mathcal{M}_l^\infty(h) + |\mathbb{E}\widehat{G}_h - G|_{2,\infty}. \tag{3.23}$$

It then remains to control the term $\widehat{\mathrm{BV}}(h)$. We have :

$$\begin{aligned}\widehat{\mathrm{BV}}(h) - \mathcal{M}_l^\infty(h) &\leq \sup_{h,h'} \left\{|\widehat{G}_{h,h'} - \mathbb{E}\widehat{G}_{h,h'}|_{2,\infty} + |\widehat{G}_{h'} - \mathbb{E}\widehat{G}_{h'}|_{2,\infty} - \mathcal{M}_l(h,h')\right\} \\ &\quad + \sup_{h'} |\mathbb{E}\widehat{G}_{h,h'} - \mathbb{E}\widehat{G}_{h'}|_{2,\infty} = \zeta + \sup_{h'} |\mathbb{E}\widehat{G}_{h,h'} - \mathbb{E}\widehat{G}_{h'}|_{2,\infty}.\end{aligned}$$

The oracle inequality follows directly from (3.23), Definition 7 and the definition of ζ. ∎

3.2.5 Adaptive fast rates in anisotropic noisy clustering

We have at our disposal a family of kernel ERM $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}\}$ defined in (3.4) with associated kernel empirical risk $\widehat{R}_h(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{c}, \cdot) * \widetilde{\mathcal{K}}_h(Z_i - \cdot)$, with $\widetilde{\mathcal{K}}_h$ a deconvolution kernel. In this paragraph, we propose to apply the selection rule (3.19) to choose the bandwidth $h \in \mathcal{H}$ in an anisotropic framework. For any $h \in \mathcal{H}$, the G-empirical risk vector is defined as :

$$\widehat{G}_h(\mathbf{c}) := \left(\frac{1}{n} \sum_{i=1}^n 2 \int_{V_j(\mathbf{c})} (x_u - c_{uj}) \widetilde{\mathcal{K}}_h(Z_i - x) dx \right)_{u=1,\dots,d, j=1,\dots,k} \in \mathbb{R}^{dk}, \ \forall \mathbf{c} \in \mathbb{R}^{dk},$$

where for any $j = 1, \dots, k$, $V_j(\mathbf{c}) := \{x \in [0,1]^d : \arg\min_{a=1,\dots,k} |x - c_a|_2 = j\}$ is the Voronoï cells associated to $\mathbf{c}$, and x_u denotes the u^{th} coordinate of $x \in \mathbb{R}^d$. Note that $\widehat{G}_h(\widehat{\mathbf{c}}_h) = (0, \dots, 0)^\top$ by smoothness. The construction of the rule follows exactly the general case of Section 3.2.4, which is based on the introduction of an auxiliary G-empirical risk. For any couple of bandwidths $(h, h') \in \mathcal{H}^2$, the auxiliary G-empirical risk is defined as :

$$\widehat{G}_{h,h'}(\mathbf{c}) := \left(\frac{1}{n} \sum_{i=1}^n 2 \int_{V_j(\mathbf{c})} (x_u - c_{uj}) \widetilde{\mathcal{K}}_{h,h'}(Z_i - x) dx \right)_{u=1,\dots,d, j=1,\dots,k} \in \mathbb{R}^{dk}, \ \forall \mathbf{c} \in \mathbb{R}^{dk},$$

where $\widetilde{K}_{h,h'} = \widetilde{\mathcal{K}_h * \mathcal{K}_{h'}}$ is the auxiliary deconvolution kernel as in Comte and Lacour [51].

The statement of the oracle inequality is based on the computation of a majorant function. For this purpose, we need to consider a kernel $\mathcal{K}_h$ of order $r \in \mathbb{N}^d$ satisfying the *Kernel assumption* (see Chapter 2), for some $S = (S_1, \dots, S_d) \in \mathbb{R}_+^d$. Additionally, we need an assumption on the noise distribution η. We suppose in the sequel that $\mathbf{NA}(\rho, \beta)$ holds, for some $\beta = (\beta_1, \dots, \beta_d) \in (0, \infty)^d$ and some positive constant ρ.

We are now ready to compute the majorant function in our context. Let $\mathcal{H} := [h_-, h^+]^d$ be the bandwidth set such that $0 < h_- < h^+ < 1$,

$$h_- := \left(\frac{\log^6(n)}{n} \right)^{1/\max(2, 2\sum_{j=1}^d \beta_j)} \quad \text{and } h^+ := \big(1/\log(n)\big)^{1/(2s^+)}, \tag{3.24}$$

for some $s^+ > 1$.

Lemma 8. *Assume the Kernel assumption and* $\mathbf{NA}(\rho, \beta)$ *hold for some* $\rho > 0$ *and some* $\beta \in \mathbb{R}_+^d$*. Let* $a \in (0,1)$ *and consider* $\mathcal{H}_a := \{(h_-, \dots, h_-)\} \cup \{h \in \mathcal{H} : \forall j = 1, \dots, d \, \exists m_j \in \mathbb{N} : h_j = h^+ a^{m_j}\}$ *an exponential net of* $\mathcal{H} = [h_-, h^+]^d$*, such that* $|\mathcal{H}_a| \leq n$*. For any integer* $l > 0$*, let us introduce the function* $\mathcal{M}_l^{\mathrm{k}} : \mathcal{H}^2 \to \mathbb{R}_+$ *defined as :*

$$\mathcal{M}_l^{\mathrm{k}}(h, h') := b_1' \sqrt{kd} \left(\frac{\Pi_{i=1}^d h_i^{-\beta_i}}{\sqrt{n}} + \frac{\Pi_{i=1}^d (h_i \vee h_i')^{-\beta_i}}{\sqrt{n}} \right),$$

where $b_1' > 0$*. Then, for* n *sufficiently large, the function* $\mathcal{M}_l^{\mathrm{k}}$ *is a majorant, i.e.*

$$\mathbb{P}\left(\sup_{h,h' \in \mathcal{H}_a} \left\{ |\widehat{G}_{h,h'} - \mathbb{E}\widehat{G}_{h,h'}|_{2,\infty} + |\widehat{G}_{h'} - \mathbb{E}\widehat{G}_{h'}|_{2,\infty} - \mathcal{M}_l^{\mathrm{k}}(h,h') \right\}_+ > 0 \right) \leq n^{-l},$$

where $\mathbb{E}$ *denotes the expectation w.r.t. the sample and* $|T|_{2,\infty} = \sup_{\mathbf{c} \in [0,1]^{dk}} |T(\mathbf{c})|_2$ *for* $T : \mathbb{R}^{dk} \to \mathbb{R}^{dk}$ *with* $|\cdot|_2$ *the Euclidean norm on* $\mathbb{R}^{dk}$*.*

The proof is based on a chaining argument and a Talagrand's inequality. This lemma is the cornerstone of the oracle inequality below, and gives the order of the variance term in such a problem.

We are now ready to define the EGC selection rule in noisy clustering as :

$$\widehat{h} = \arg\min_{h \in \mathcal{H}_a} \left\{ \sup_{h' \in \mathcal{H}_a} \left\{ |\widehat{G}_{h,h'} - \widehat{G}_{h'}|_{2,\infty} - \mathcal{M}_l^{\mathrm{k}}(h,h') \right\} + \mathcal{M}_l^{\mathrm{k},\infty}(h) \right\}, \tag{3.25}$$

where $\mathcal{M}_l^{\mathrm{k},\infty}(h) := \sup_{h' \in \mathcal{H}_a} \mathcal{M}_l^{\mathrm{k}}(h', h)$. The next theorem gives a control of the G-excess risk of the kernel ERM $\widehat{\mathbf{c}}_{\widehat{h}}$.

Corollary 4. *Assume* $\mathbf{NA}(\rho, \beta)$ *hold for some* $\rho > 0$ *and some* $\beta \in \mathbb{R}_+^d$*. Then, for* n *large enough, With probability* $1 - n^{-l}$*, it holds :*

$$|G(\widehat{\mathbf{c}}_{\widehat{h}}, \mathbf{c}^\star)|_2 \leq 3 \inf_{h \in \mathcal{H}_a} \left\{ B^{\mathrm{k}}(h) + \mathcal{M}_l^{\mathrm{k},\infty}(h) \right\},$$

where $B^{\mathrm{k}} : \mathcal{H} \to \mathbb{R}_+$ *is a bias function defined as :*

$$B^{\mathrm{k}}(h) := 2\sqrt{k}\,(1 \vee |\mathcal{F}[\mathcal{K}]|_\infty)\, |\mathcal{K}_h * f - f|_2, \quad \forall h \in \mathcal{H}.$$

The proof of Theorem 4 is an application of Theorem 11 gathering with Lemma 8. Note that the infimum in the RHS is restricted over the net $\mathcal{H}_a$. However, as shown in Theorem 12 below, this is sufficient to obtain adaptive optimal fast rates.

As mentioned in the previous section, we can deduce fast rates for the excess risk as an important theorem. For this purpose, we need an additional assumption on the regularity of the density f to control the bias function in Theorem 4. This regularity is expressed here in terms of anisotropic Nikol'skii space.

Definition 8 (Anisotropic Nikol'skii Space)**.** *Let* $s = (s_1, s_2, \ldots, s_d) \in \mathbb{R}_+^d$*,* $q \geq 1$ *and* $L > 0$ *be fixed. We say that* $f : [0,1]^d \to [-L, L]$ *belongs to the anisotropic Nikol'skii space* $\mathcal{N}_q(s, L)$ *of functions if for all* $j = 1, ..., d$*,* $z \in \mathbb{R}$ *and for all* $x \in (0,1]^d$ *:*

$$\left(\int \left| D_j^{\lfloor s_j \rfloor} f(x_1, \ldots, x_j + z, \ldots, x_d) - D_j^{\lfloor s_j \rfloor} f(x_1, \ldots, x_j, \ldots, x_d) \right|^q dx \right)^{1/q} \leq L|z|^{s_j - \lfloor s_j \rfloor},$$

and $\|D_j^l f\|_q \leq L$*,* $\forall l = 0, \ldots, \lfloor s_j \rfloor$*, where* $D_j^l f$ *denotes the* l*-th order partial derivative of* f *w.r.t. the variable* x_j *and* $\lfloor s_j \rfloor$ *is the largest integer strictly less than* s_j*.*

The Nikol'skii spaces have been considered in approximation theory by Nikol'skii (see Nikol'skii [1975] for example). We also refer to Goldenshluger and Lepski [2011], Kerkyacharian, Lepski, and Picard [2001] where the problem of adaptive estimation over a scale s has been treated for the Gaussian white noise model and for density estimation, respectively.

In the sequel, we assume that the multivariate density f of the law P_X belongs to the anisotropic Nikol'skii class $\mathcal{N}_2(s, L)$, for some $s \in \mathbb{R}_+^d$ and some $L > 0$. It means that the density f has possible different regularities in all directions.

Theorem 12. *Assume the Kernel assumption and* $\mathbf{NA}(\rho, \beta)$ *hold for some* $\rho > 0$ *and some* $\beta \in \mathbb{R}_+^d$*. Assume the Hessian matrix of* R *is positive definite for any* $\mathbf{c}^\star \in \mathcal{M}$*. Then, for any* $s \in (0, s^+)^d$*,* $L > 0$ *:*

$$\limsup_{n \to \infty} n^{1/(1+\sum_{j=1}^d \beta_j / s_j)} \sup_{f \in \mathcal{N}_2(s,L)} \mathbb{E}\left[R(\widehat{\mathbf{c}}_{\widehat{h}}) - R(\mathbf{c}^\star) \right] < \infty,$$

where $\widehat{h}$ *is chosen in* (3.25)*.*

This theorem uses Corollary 4 and Lemma 7, gathering with the consistency of the family of kernel ERM $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}\}$. In this respect, the definitions of h_- and h^+ in (3.24), gathering with the continuity of the density f, imply the consistency of our family.

This result gives adaptive fast rates for the excess risk of $\widehat{\mathbf{c}}_{\widehat{h}}$. It improves the result stated in Section 3.1.2 with ERC (see Theorem 9) for two main reasons. First of all, the selection rule allows the extension to the anisotropic case. Besides, there is no logarithmic term in the adaptive rate. The localization technique used in the previous result seems the main obstacle to avoid the extra-log term. The G-excess risk approach avoids the localization technique and therefore the extra-log term in the adaptive fast rates. The result of Theorem 12 also extend the result to Nikol'skii spaces instead of Hölder spaces.

3.2.6 Anisotropic adaptive minimax rates for nonlinear estimators

In this subsection, we apply the result of Theorem 11 to state minimax adaptive rates of convergence for nonlinear estimators in the framework of local M-estimation. As in noisy clustering, the previous oracle inequality for the G-excess risk will give us adaptive minimax results for both pointwise and global estimation.

Let us specify the model beforehand. For some $n \in \mathbb{N}^*$, we observe a training set $\mathcal{Z}_n := \{(W_i, Y_i),\ i = 1, ...n\}$ of i.i.d. pairs distributed according to the probability measure P on $[0,1]^d \times \mathbb{R}$ satisfying the set of equations :

$$Y_i = f^\star(W_i) + \xi_i, \quad i = 1, \dots, n, \tag{3.26}$$

where the noise variables $(\xi_i)_{i=1,\dots,n}$ are i.i.d. with symmetric density g_ξ w.r.t. the Lebesgue measure. We aim at estimating the target function $f^\star : [0,1]^d \to [\text{-}B, B],\ B > 0$. Moreover, we also assume that g_ξ is continuous at 0 and $g_\xi(0) > 0$. For simplicity, in the sequel, the design points $(W_i)_{i=1,\dots,n}$ are i.i.d. according to the uniform law on $[0,1]^d$ (extension to a more general design is straightforward) and we suppose that $(W_i)_{i=1,\dots,n}$ and $(\xi_i)_{i=1,\dots,n}$ are mutually independent for ease of exposition. Eventually, we restrict the estimation of $f^\star$ to the closed set $\mathcal{T} \subset [0,1]^d$ to avoid discussion on boundary effects. We will consider the point $x_0 \in \mathcal{T}$ for pointwise estimation and the $\mathbb{L}_q(\mathcal{T})$-risk for global estimation.

Next, we introduce an estimate of $f^\star(x_0)$ at any $x_0 \in \mathcal{T}$ with the local constant approach (LCA) with a fixed bandwidth. The key idea of LCA, as described for example in [167, Chapter 1], is to approximate the target function in a neighborhood of size $h \in (0,1)^d$ of a given point x_0 by a constant, which corresponds to a model of dimension $m = 1$. To deal with heavy-tailed noises, we especially employ the popular Huber loss (see Huber [1964]) defined as follows. For any scale $H > 0$ and $z \in \mathbb{R}$,

$$\rho_H(z) := \begin{cases} z^2/2 & \text{if } |z| \leq H \\ \\ H(|z| - H/2) & \text{otherwise.} \end{cases}$$

The parameter $H > 0$ selects the level of robustness of the Huber loss between the square loss (large value of H) and the absolute loss (small value of H).

Let $\mathcal{H} := [h_-, h^+]^d$ be the bandwidth set such that $0 < h_- < h^+ < 1$,

$$h_- := \frac{\log^{6/d}(n)}{n^{1/d}} \quad \text{and} \quad h^+ := \frac{1}{\log^2(n)}.$$

For any $x_0 \in \mathcal{T}$, the local estimator $\widehat{f}_h(x_0)$ of $f^\star(x_0)$ is defined as[3] :

$$\widehat{f}_h(x_0) := \underset{t \in [\text{-}B,B]}{\arg\min}\, \widehat{R}_h^{\text{loc}}(t), \quad h \in \mathcal{H},$$

where $\widehat{R}_h^{\text{loc}}(\cdot) := \frac{1}{n}\sum_{i=1}^n \rho_H(Y_i - \cdot)\ \mathcal{K}_h(W_i - x_0)$ is the local empirical risk and $\mathcal{K}_h$ is a 1-Lipschitz, nonnegative kernel of order 1. As in (3.16), the expectation of the local empirical risk has a limit denoted by $R^{\text{loc}}(\cdot) := \mathbb{E}_{Y|W=x_0}\rho_H(Y - \cdot)$ whose its unique minimizer is $f^\star(x_0)$.

To end up this chapter, we are interested in the bandwidth selection problem in the family $\{\widehat{f}_h, h \in \mathcal{H}\}$, where $\mathcal{H}$ is defined above. We want to state minimax adaptive results for both pointwise and global risks. Since Theorem 11 controls the G-excess risk of the adaptive estimator, we present the following lemma that gives rive to a control of the pointwise risk. A same inequality can be deduced with the $\mathbb{L}_q(\mathcal{T})$-norm.

Lemma 9. *Assume that* $\sup_{h\in\mathcal{H}} |\widehat{f}_h(x_0) - f^\star(x_0)| \leq \mathbb{E}\rho_H''(\xi_1)/4$. *Then, for all* $h \in \mathcal{H}$,

$$|\widehat{f}_h(x_0) - f^\star(x_0)| \leq \frac{2}{\mathbb{E}\rho_H''(\xi_1)} \left| G^{\text{loc}}\big(\widehat{f}_h(x_0)\big) - G^{\text{loc}}\big(f^\star(x_0)\big) \right|,$$

where G^{loc} *(and resp.* ρ_H''*) denotes the derivative of* R^{loc} *(resp. the second derivative of* ρ_H*).*

[3] We use in the sequel $\widehat{f}_h(\cdot)$ for the estimator of $f^\star$ in (3.26). This estimator is NOT a kernel estimator $\hat{f}_h(\cdot)$ as before, but a minimizer of a kernel empirical risk.

The assumption $\sup_{h\in\mathcal{H}} |\widehat{f}_h(x_0) - f^\star(x_0)| \le \mathbb{E}\rho''_H(\xi_1)/4$ is necessary to use the theory of differential calculus and can be satisfied by using the consistency of $\widehat{f}_h$. In this direction, the definitions of h_- and h^+ above imply the consistency of all estimators $\widehat{f}_h, h \in \mathcal{H}$ (see Theorem 1 in Chichignoud and Lederer [2013] for further details). This lemma allows us to link the local G-excess risk and the pointwise semi-norm.

The selection rule in pointwise estimation

To compute the selection procedure in pointwise estimation, we define the G-empirical risk as :

$$\widehat{G}_h^{\mathrm{loc}}(t) := \frac{\partial \widehat{R}_h^{\mathrm{loc}}}{\partial t}(t) = -\frac{1}{n}\sum_{i=1}^n \rho'_H\big(Y_i - t\big)\, \mathcal{K}_h(W_i - x_0). \tag{3.27}$$

For any couple of bandwidths $(h, h') \in \mathcal{H}^2$, we introduce the auxiliary G-empirical risk as :

$$\widehat{G}_{h,h'}^{\mathrm{loc}}(t) := -\frac{1}{n}\sum_{i=1}^n \rho'_H\big(Y_i - t\big)\, \mathcal{K}_{h,h'}(W_i - x_0),$$

where $\mathcal{K}_{h,h'} := \mathcal{K}_h * \mathcal{K}_{h'}$ as above.

To apply the results of Section 3.2.4, we need to compute optimal majorants of the associated empirical processes. The construction of such bounds for the pointwise case has already deserved some interests. For any integer $l \in \mathbb{N}^*$, let us introduce the function $\mathcal{M}_l^{\mathrm{loc}} : \mathcal{H}^2 \to \mathbb{R}_+$ defined as :

$$\mathcal{M}_l^{\mathrm{loc}}(h, h') := C_0 \|\mathcal{K}\|_2 \sqrt{\mathbb{E}[\rho'_H(\xi_1)]^2}\left(\sqrt{\frac{l\log(n)}{n\prod_{i=1}^d h_i \vee h'_i}} + \sqrt{\frac{l\log(n)}{n\prod_{i=1}^d h'_i}}\right),$$

where $C_0 > 0$ is an absolute constant which does not depend on the model.
Let $\mathcal{H}_a := \{(h_-, \dots, h_-)\} \cup \{h \in \mathcal{H} \,:\, \forall j = 1, \dots, d\, \exists m_j \in \mathbb{N} \,:\, h_j = h^+ a^{m_j}\},\ a \in (0,1)$, be an exponential net of $\mathcal{H} = [h_-, h^+]^d$, such that $|\mathcal{H}_a| \le n$. Then, for any $l > 0$, the function $\mathcal{M}_l^{\mathrm{loc}}(\cdot,\cdot)$ is a majorant, i.e. :

$$\mathbb{P}\left(\sup_{h,h'\in\mathcal{H}_a}\left\{|\widehat{G}_{h,h'}^{\mathrm{loc}} - \mathbb{E}\widehat{G}_{h,h'}^{\mathrm{loc}}|_\infty + |\widehat{G}_{h'}^{\mathrm{loc}} - \mathbb{E}\widehat{G}_{h'}^{\mathrm{loc}}|_\infty - \mathcal{M}_l^{\mathrm{loc}}(h,h')\right\}_+ > 0\right) \le n^{-l},$$

We notice that, unlike Definition 7, $|\cdot|_{2,\infty}$ is replaced by $|\cdot|_\infty$ since the G-empirical risk (3.27) is unidimensional.

Eventually, we introduce the data-driven bandwidth following the schema of the selection rule in Section 3.2.4 :

$$\widehat{h}^{\mathrm{loc}} := \arg\min_{h\in\mathcal{H}_a}\left\{\sup_{h'\in\mathcal{H}_a}\left\{|\widehat{G}_{h,h'}^{\mathrm{loc}} - \widehat{G}_{h'}^{\mathrm{loc}}|_\infty - \mathcal{M}_l^{\mathrm{loc}}(h,h')\right\} + 2\mathcal{M}_l^{\mathrm{loc},\infty}(h)\right\}, \tag{3.28}$$

where $\mathcal{M}_l^{\mathrm{loc},\infty}(h) := \sup_{h'\in\mathcal{H}_a} \mathcal{M}_l^{\mathrm{loc}}(h', h)$. We are now ready to give the oracle inequality for the pointwise risk :

Corollary 5. *Consider the model* (3.26) *and assume that n is great enough. Then, for any $l > 0$, with probability $1 - n^{-l}$, we have :*

$$|\widehat{f}_{\widehat{h}^{\mathrm{loc}}}(x_0) - f^\star(x_0)| \le \frac{6}{\mathbb{E}\rho''_H(\xi_1)} \inf_{h\in\mathcal{H}_a}\left\{B^{\mathrm{loc}}(h) + 2\mathcal{M}_l^{\mathrm{loc},\infty}(h)\right\},$$

where $B^{\mathrm{loc}}(h)$ denotes the bias term $B^{\mathrm{loc}}(h) := \int \mathcal{K}_h(x-y)\,|f^\star(x) - f^\star(y)|\,dx$.

The proof is a direct application of Theorem 11 and Lemma 9, since $G^{\mathrm{loc}}(f^\star(x_0)) = 0$ and

$$\sup_{h'\in\mathcal{H}} \mathcal{M}_l^{\mathrm{loc}}(h', h) = \mathcal{M}_l^{\mathrm{loc},\infty}(h).$$

Note that the infimum in the RHS of Theorem 5 is restricted to the net $\mathcal{H}_a$. However, as shown in Theorem 13 below, this is sufficient to obtain minimax adaptive results.

Chichignoud and Lederer [2013] have shown that the variance of local M-estimators is of order $\mathbb{E}[\rho'_H(\xi_1)]^2/n(\mathbb{E}\rho''_H(\xi_1))^2$. Therefore, their Lepski-type procedure depends on this quantity. Here, we obtain the same result without the dependency on the parameter $\mathbb{E}\rho''_H(\xi_1)$ - which corresponds to $\lambda_{\min}$ in the general setting - thanks to the gradient approach. The selection rule is therefore robust w.r.t. to the fluctuations of this parameter, in particular when H is small (median estimator).

Now, we focus on the minimax issue for pointwise estimation by giving the following theorem :

Theorem 13. *For any $s \in (0,1]^d$, any $L > 0$ and any $q \geq 1$, it holds for all $x_0 \in \mathcal{T}$:*

$$\limsup_{n\to\infty} (n/\log(n))^{q\bar{s}/(2\bar{s}+1)} \sup_{f^\star \in \Sigma(s,L)} \mathbb{E}\left|\widehat{f}_{\widehat{h}^{\mathrm{loc}}}(x_0) - f^\star(x_0)\right|^q < \infty,$$

where $\bar{s} := \left(\sum_{j=1}^d s_j^{\text{-}1}\right)^{\text{-}1}$ denotes the harmonic average and $\Sigma(s,L)$ denotes the anisotropic Hölder class of Definition 5 (see Chapter 2).

The proposed estimator $\widehat{f}_{\widehat{h}}$ is then adaptive minimax over anisotropic Hölder spaces in pointwise estimation. The minimax optimality of this rate (with the $\log(n)$ factor) has been stated by Klutchnikoff [2005] in the white noise model for pointwise estimation (see also Goldenshluger and Lepski [2008]). We did not study the case of locally polynomial functions, which is further complicated to study in nonparametric regression. In this case, we could consider smoother functions $f^\star \in \Sigma(s,L)$, with $s \in (0,s^+)^d$, $s^+ > 1$.

The selection rule in global estimation

The aim of this paragraph is to derive adaptive minimax results for $\widehat{f}_h$ in $\mathbb{L}_q$-risk. To this end, we need to modify the selection rule (3.28) including a global ($\mathbb{L}_q$-norm) comparison of G-empirical risks. For this purpose, for all $t \in \mathbb{R}$, we denote the G-empirical risks at a given point $x_0 \in \mathcal{T}$ as :

$$\widehat{G}_h^{\mathrm{glo}}(t,x_0) = -\frac{1}{n}\sum_{i=1}^n \rho'_H\big(Y_i - t\big)\,\mathcal{K}_h(W_i - x_0) \quad\text{and}\quad \widehat{G}_{h,h'}^{\mathrm{glo}}(t,x_0) = -\frac{1}{n}\sum_{i=1}^n \rho'_H\big(Y_i - t\big)\,\mathcal{K}_{h,h'}(W_i - x_0),$$

where the dependence in x_0 is explicitly written. We then define, for $q \in [1,\infty[$ and for any function $\omega : \mathbb{R}\times\mathcal{T} \to \mathbb{R}$, the $\mathbb{L}_q$-norm and $\mathbb{L}_{q,\infty}$-semi-norm :

$$\|\omega(t,\cdot)\|_q := \left(\int_{\mathcal{T}} |\omega(t,x)|^q dx\right)^{1/q} \quad\text{and}\quad \|\omega\|_{q,\infty} := \sup_{t\in[-B,B]} \|\omega(t,\cdot)\|_q.$$

The construction of majorants is based on uniform bounds for $\mathbb{L}_q$-norms of empirical processes. This topic has been recently investigated in Goldenshluger and Lepski [2011]. For any $l \in \mathbb{N}^*$, let us introduce the function $\Gamma_{l,q} : \mathcal{H} \to \mathbb{R}_+$ defined as :

$$\Gamma_{l,q}(h) := C_q\|\rho'_H\|_\infty\sqrt{1+l}\times\begin{cases} 4\|\mathcal{K}\|_q(n\Pi_h)^{-(q-1)/q} & \text{if } q\in[1,2[, \\ \frac{30q}{\log(q)}(\|\mathcal{K}\|_2 \vee \|\mathcal{K}\|_q)(n\Pi_h)^{-1/2} & \text{if } \ q\in[2,\infty[, \end{cases}$$

where $\Pi_h = \prod_{j=1}^d h_j$ and $C_q > 0$ is an absolute constant which does not depend on n. Then, for any $l > 0$, the function $\mathcal{M}_{l,q}^{\mathrm{glo}}(h,h') := \Gamma_{l,q}^{\mathrm{glo}}(h\vee h') + \Gamma_{l,q}^{\mathrm{glo}}(h')$ is a majorant, i.e. :

$$\mathbb{P}\left(\sup_{h,h'\in\mathcal{H}}\left\{\|\widehat{G}_{h,h'}^{\mathrm{glo}} - \mathbb{E}\widehat{G}_{h,h'}^{\mathrm{glo}}\|_{q,\infty} + \|\widehat{G}_{h'}^{\mathrm{glo}} - \mathbb{E}\widehat{G}_{h'}^{\mathrm{glo}}\|_{q,\infty} - \mathcal{M}_{l,q}^{\mathrm{glo}}(h,h')\right\}_+ > 0\right) \leq n^{-l},$$

where the constant C_q can be explicitly given.

We finally select the bandwidth according to the EGC rule in Section 3.2.4 :

$$\widehat{h}_q^{\mathrm{glo}} := \arg\min_{h\in\mathcal{H}}\left\{\sup_{h'\in\mathcal{H}}\left\{\|\widehat{G}_{h,h'}^{\mathrm{glo}} - \widehat{G}_{h'}^{\mathrm{glo}}\|_{q,\infty} - \mathcal{M}_{l,q}^{\mathrm{glo}}(h,h')\right\} + 2\Gamma_{l,q}(h)\right\}.$$

Corollary 6. *Consider the model* (3.26) *and assume that n is great enough. For any $l > 0$, we then have with probability $1 - n^{-l}$:*

$$\|\widehat{f}_{\widehat{h}_q^{\mathrm{glo}}} - f^\star\|_q \leq \frac{6}{\mathbb{E}\rho''_H(\xi_1)} \inf_{h \in \mathcal{H}} \left\{ B_q^{\mathrm{glo}}(h) + 2\Gamma_{l,q}^{\mathrm{glo}}(h) \right\},$$

where $B_q^{\mathrm{glo}}(h) := \left\| \int \mathcal{K}_h(x - \cdot) \left| f^\star(x) - f^\star(\cdot) \right| dx \right\|_q$ is called the global bias term.

We notice that there is no restriction about the infimum over $\mathcal{H}$ - compared to the local oracle inequality - which is due to the construction of majorant. The proof is based on the same scheme as the proof of Theorem 11, by adding the $\mathbb{L}_q$-norm. Gathering with a global version of Lemma 9 (i.e. a control of the $\mathbb{L}_q$-norm instead of the pointwise semi-norm), we get the result.

The above choice of the bandwidth leads to the estimator $\widehat{f}_{\widehat{h}_q^{\mathrm{glo}}}$ with the following adaptive minimax properties for the $\mathbb{L}_q$-risk over anisotropic Nikol'skii spaces (see Definition 8 in Section 3.2.5).

Theorem 14. *For any $s \in (0,1]^d$, any $L > 0$ and any $q \geq 1$, it holds :*

$$\limsup_{n \to \infty} \psi_{n,q}^{-1}(s) \sup_{f^\star \in \mathcal{N}_{q,d}(s,L)} \mathbb{E}\|\widehat{f}_{\widehat{h}_q^{\mathrm{glo}}} - f^\star\|_q^q < \infty$$

where $\bar{s} := \left(\sum_{j=1}^d s_j^{-1} \right)^{-1}$ denotes the harmonic average and

$$\psi_{n,q}(s) := \begin{cases} (1/n)^{q(q-1)\bar{s}/(q\bar{s}+q-1)} & \text{if } q \in [1,2[, \\ \\ (1/n)^{q\bar{s}/(2\bar{s}+1)} & \text{if } \ q \geq 2. \end{cases}$$

We refer to Has'minskii and Ibragimov [1990], Has'minskii and Ibragimov [1981] for the minimax optimality of these rates over Nikol'skii spaces. The proposed estimate $\widehat{f}_{\widehat{h}_q^{\mathrm{glo}}}$ is then adaptive minimax. To the best of our knowledge, the minimax adaptivity over anisotropic Nikol'skii spaces has never been done in regression with possible heavy-tailed noises. As in pointwise estimation, this result could be extended to the case of local polynomial functions of order $k \geq 1$.

3.2.7 Discussion

This section deals with the bandwidth selection problem in kernel empirical risk minimization. We propose a new criterion called the gradient excess risk (3.10), which allows us to derive optimal fast rates of convergence for the excess risk as well as adaptive minimax rates for global and pointwise risks.

One of the key messages we would like to highlight is the following : if we consider smooth loss functions and a family of consistent ERM, fast rates of convergence are automatically reached provided that the Hessian matrix of the risk function is positive definite. This statement is based on the key Lemma 5 where the square root of the excess risk is controlled by the G-excess risk.

From an adaptive point of view, another look at Lemma 5 can be done. In the RHS of Lemma 5, the G-excess risk is multiplied by the constant $\lambda_{\min}^{-1}$, i.e. the smallest eigenvalue of the Hessian matrix at $\theta^\star$. This parameter is also involved in the margin assumption (see Lemma 6). As a result, our selection rule does not depend on this parameter since the margin assumption is not required to obtain slow rates for the G-excess risk. This fact partially solves an issue highlighted by Massart [126, Section 8.5.2], in the model selection framework :

> *"It is indeed a really hard work in this context to design margin adaptive penalties. Of course recent works on the topic, involving local Rademacher penalties for instance, provide at least some theoretical solution to the problem but still if one carefully looks at the penalties which are proposed in these works, they systematically involve constants which are typically unknown. In some cases, these constants are absolute constants which should nevertheless considered as unknown just because the numerical values coming from the theory are obviously over pessimistic. In some other cases, it is even worse since they also depend on nuisance parameters related to the unknown distribution."*

We can also mention the work of Koltchinskii [100], who has studied the general margin assumption. In this context, a "link function" $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ describes the relationship between the excess risk and the variance term, i.e.

$$\varphi\left(\sqrt{\mathbb{E}_{P_{\mathcal{X}}}\left[\ell(\mathcal{X},\theta)-\ell(\mathcal{X},\theta^\star)\right]^2}\right) \leq R(\theta)-R(\theta^\star),$$

for all θ belongs to a ball of $\theta^\star$. In our context, with smooth loss functions, the link function corresponds to the square function : $\varphi(x) = Cx^2$, $\forall x \in \mathbb{R}_+$ with $C = \lambda_{\min}/(3\kappa_1)$ (see Lemma 6). Koltchinskii [100, Section 6.3] has highlighted the issue of the adaptivity w.r.t. the link function as follows :

> *"It happens that the link function is involved in a rather natural way in the construction of complexity penalties that provide optimal convergence rates in many problems. Since the link function is generally distribution dependent, the development of adaptive penalization methods of model selection is a challenge, for instance, in classification setting."*

3.3 Computation of ERC and EGC [L12]

In this chapter, adaptive fast rates of convergence have been proved for two bandwith selection methods. The first one, called ERC, allows to obtain good theoretical guarantees in the isotropic case. The second one, called EGC, allows to consider the anisotropic framework. These methods are based on Lepski's heuristic (see Lepski [1990]) and consists in comparing empirical criteria (such as empirical risk in ERC or empirical gradient for EGC) for different values of the bandwidth. Whereas the origin of Lepski's method is mainly theoretical, practical issues have also deserved some attentions in the last decade. The main contribution to this field is perhaps the intersection of confidence intervals (ICI) rule (see Katkovnik [1999]), which computes the isotropic Lepski's principle in a 1-dimensional grid of increasing bandwidths. This algorithm is at the core of many applications in image denoising, where data-driven selection rule are of practical interest (see Kervrann and Boulanger [2006], Astola, Egiazarian, Foi, and Katkovnik [2010] and references therein). For the anisotropic case, Comte and Lacour [2013] also computes the anisotropic GL method in a deconvolution setting. In this section, we investigate the computation of the two selection methods proposed in Section 3.1-3.2 in the framework of clustering with errors-in-variables.

3.3.1 Model and notations

In this section, we are interested in clustering with errors-in-variables. Let us consider a noisy sample :

$$Z_i = X_i + \epsilon_i, i = 1, \dots, n, \tag{3.29}$$

where as before X_i, $i = 1, \dots, n$ are i.i.d. with density f with respect to the Lebesgue measure and ϵ_i, $i = 1, \dots, n$ are i.i.d. with density η. Given some integer $k \geq 2$, we want to summarize the dataset X_i, $i = 1, \dots, n$ with the sequence of observations Z_i, $i = 1, \dots, n$. As a rule, we use a density deconvolution estimator :

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^n \widetilde{\mathcal{K}}_h\left(Z_i - x\right),$$

where $h \in \mathbb{R}_+^d$ is a bandwidth parameter. In the sequel, we are interested in the data-driven choice of h. Given a kernel deconvolution estimator $\hat{f}_h$, we consider the family of noisy k-means estimators $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}\}$, where for a given bandwidth h, $\widehat{\mathbf{c}}_h$ is defined as the minimizer of the deconvolution empirical risk $\widehat{R}_h(\cdot)$ defined in (3.5).

The empirical risk $\widehat{R}_h(\cdot)$ is of first interest in this section. This quantity will be evaluated for different values of $h \in \mathcal{H}$, where $\mathcal{H} \subset \mathbb{R}_+$ is a given grid of bandwidth. More precisely, the computation of $\widehat{R}_h(\cdot)$ for increasing values of h will be at the core of the ICI rule defined in Section 3.3.2. However, as it was shown in Section 3.2, this empirical risk is not suitable in the anisotropic case. As a result, we introduce in the sequel a second bandwidth selection based on the computations of the gradient of $\widehat{R}_h(\cdot)$. For any given $h \in \mathbb{R}_+^d$, we also defined the empirical gradient as :

$$\widehat{G}_h(\mathbf{c}) = \left(\frac{1}{n} \sum_{i=1}^n 2 \int_{V_j(\mathbf{c})} (x_u - c_{uj}) \widetilde{\mathcal{K}}_h(Z_i - x) dx \right)_{u=1,\dots,d, j=1,\dots,k} \in \mathbb{R}^{dk}, \ \forall \mathbf{c} \in \mathbb{R}^{dk}, \tag{3.30}$$

where for any $j = 1, \dots, k$, $V_j(\mathbf{c}) := \{x \in [0,1]^d : \arg\min_{a=1,\dots,k} |x - c_a|_2 = j\}$ is the Voronoï cells associated to $\mathbf{c}$, and x_u denotes the u^{th} coordinate of $x \in \mathbb{R}^d$. In the sequel, we suggest to compare (3.30) at different values of h in order to construct an data-driven bandwidth $\widehat{h}$ in the anisotropic framework.

Note that to construct the family of estimators $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}\}$, we use an alteration of the popular k-means algorithm of Hartigan [1975]. At each iteration, a deconvolution kernel function is involved in the Newton optimization. Unfortunately, the minimization problem is not convex and we can only compute a local minimizer. As a result, the solution depends strongly on the initialization step in the algorithm and affects significantly the problem of bandwidth selection. At the light of Section 3.1, ERC rule compares empirical risks $\widehat{R}_h(\cdot)$ at given global minimizers $\widehat{\mathbf{c}}_h$, which is not achievable in practice. In Section 3.2, EGC rule is introduced as a non-convex optimization problem related with the minimization of (3.30). With these considerations in mind, we expect that, up to some optimization intrinsic difficulties, computations of ERC and EGC can lead to efficient performances, at least in comparison with standard k-means. In this direction, multiple initializations could be proposed for both the noisy k-means algorithm and the choice of the bandwidth (see Open Problem 10).

In the isotropic case, we can consider a one-dimensional grid $\mathcal{H} \subset \mathbb{R}$ made of L values. We denote it as $\mathcal{H}_{\text{iso}}$ in the sequel. Equipped with this grid, we use a sequential procedure based on the ICI rule to deal with the isotropic choice of the bandwidth. Loosely speaking, for increasing values of bandwidths $h \in \mathcal{H}_{\text{iso}}$, we construct an intersection of confidence intervals and stops when this intersection is the empty set. In the anisotropic case, we restrict the study to the two-dimensional case for computational issues ($d = 2$ in (3.29)). We consider a two-dimensional bandwidth (h_1, h_2) and consider a set $\mathcal{H}_{\text{aniso}}$ of $L \times L$ values. Given this two-dimensional grid, we minimize an estimate of the bias-variance decomposition of the gradient excess risk. This estimation is computed thanks to (3.30) and the introduction of an auxiliary empirical gradient defined below.

3.3.2 ICI rule for ERC

The ICI method is a now popular bandwidth selection method. It was proposed by Katkovnik [1999] as an alteration to the theoretical Lepski's method. The implementation is very simple and does not need the computation of all the estimators in the family, in comparison to the Lepski's method. It has been successfully applied in various areas, such as image processing (see Astola, Egiazarian, and Katkovnik [2002], Astola, Egiazarian, Foi, and Katkovnik [2010]). In our case, we want to use an ICI-based method to implement the ERC method.

In Section 3.1, the ERC selection rule allows a theoretical well justified method to design noisy k-means with adaptive properties. The selected bandwidth does not depend on the regularity of the density f in (3.29). The data-driven bandwidth chosen with ERC is given by :

$$\hat{h} = \max \left\{ h \in \mathcal{H}_{\text{iso}} \ : \ \widehat{R}_{h'}(\hat{\mathbf{c}}_h) - \widehat{R}_{h'}(\hat{\mathbf{c}}_{h'}) \leq 3\delta_{h'}, \, \forall h' \leq h \right\}. \tag{3.31}$$

As discussed later on, the principal motivation to introduce ERC in to compare empirical risks instead of estimators. Then, in order to apply the ICI rule to (3.31), we choose to replace intervals centered at pointwise estimators (see Katkovnik [1999]) by intervals centered at empirical risks $\widehat{R}_h(\widehat{\mathbf{c}}_h)$. This motivates the introduction of a sequence of intervals $(\mathcal{D}_k)_{k=1}^L$ such that :

$$\mathcal{D}_k = \left[\widehat{\mathcal{R}}_k - C_{\text{iso}} \frac{h_k^{-2\beta} \log(n)}{n} ; \widehat{\mathcal{R}}_k + C_{\text{iso}} \frac{h_k^{-2\beta} \log(n)}{n} \right[, \ \forall k = 1, \dots, L, \tag{3.32}$$

where in (3.32), $C_{\text{iso}} > 0$ and for any bandwidth $h_k \in \mathcal{H}_{\text{iso}}$, $\widehat{\mathcal{R}}_k := \widehat{R}_{h_k}(\widehat{c}_{h_k})$. Then, the selected bandwidth $\widehat{h}_{\text{ICI}}$ according to ICI rule is selected according to :

$$\widehat{h}_{\text{ICI}} := \max\{h_k,\ k = 1, \dots, |\mathcal{H}_{\text{iso}}| : \mathcal{I}_k \neq \emptyset\} \text{ where } \mathcal{I}_k = \bigcap_{j=1}^{k} \mathcal{D}_k. \tag{3.33}$$

The ICI rule (3.33) can be interpreted as follows. The first interval $\mathcal{D}_1$ is constructed thanks to (3.32) with h_1. Then, the second interval $\mathcal{D}_2$ is constructed with $h_2 > h_1$ and $\mathcal{I}_2 = \mathcal{D}_1 \cap \mathcal{D}_2$ is computed. If $\mathcal{I}_2 = \emptyset$, the algorithm stops and the selected bandwidth is $\widehat{h}_{\text{ICI}} = h_1$. Otherwise, $\mathcal{D}_3$ is constructed and $\mathcal{I}_3 = \mathcal{I}_2 \cap \mathcal{D}_3$ is built. If $\mathcal{I}_3 = \emptyset$, the algorithm stops and $\widehat{h}_{\text{ICI}} = h_2$. At each iteration k, a new intersection $\mathcal{I}_k$ is obtained and we stop when the result has no point. The selected bandwidth is the maximal value of k such that $\mathcal{I}_k \neq \emptyset$. Figure 3.1 illustrates the method. It is important to notice that the chosen bandwidth made the better compromise between bias and variance of the decomposition of the excess risk. Indeed, when k increases in the algorithm, the bias increases whereas the variance decreases. Then, the lengths of the $\mathcal{I}_k$'s are decreasing whereas the centers of $\mathcal{I}_k$'s have increasing variability. As a result, we propose to stop the algorithm when the intersection of intervals $\mathcal{D}_k$ becomes the empty set.

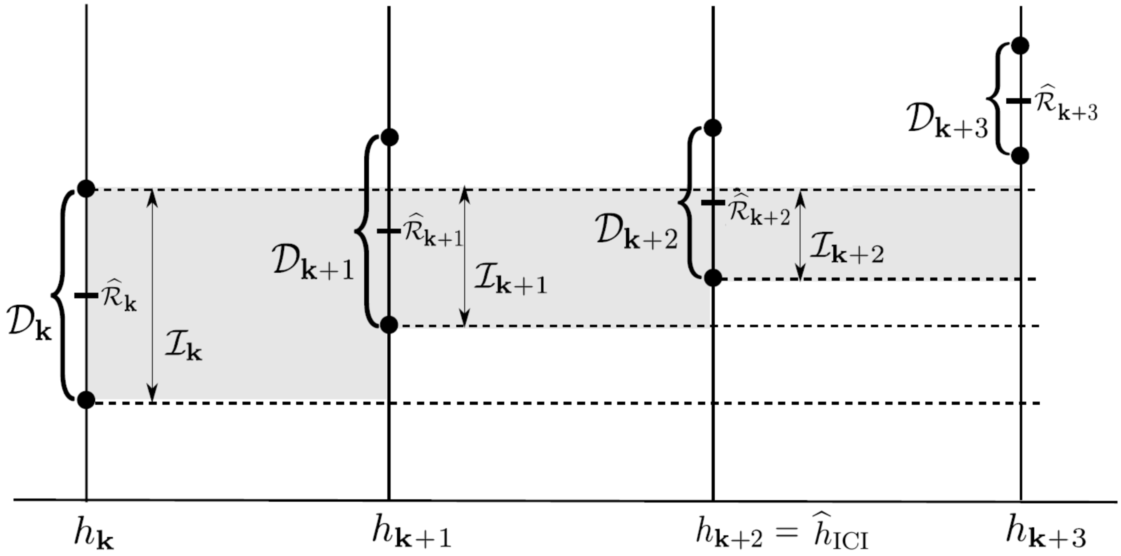

$\mathcal{D}_{\mathbf{k}}$
$\widehat{\mathcal{R}}_{\mathbf{k}}$
$\mathcal{I}_{\mathbf{k}}$
$\mathcal{D}_{\mathbf{k}+1}$
$\widehat{\mathcal{R}}_{\mathbf{k}+1}$
$\mathcal{I}_{\mathbf{k}+1}$
$\mathcal{D}_{\mathbf{k}+2}$
$\widehat{\mathcal{R}}_{\mathbf{k}+2}$
$\mathcal{I}_{\mathbf{k}+2}$
$\mathcal{D}_{\mathbf{k}+3}$
$\widehat{\mathcal{R}}_{\mathbf{k}+3}$
$h_{\mathbf{k}}$
$h_{\mathbf{k}+1}$
$h_{\mathbf{k}+2} = \widehat{h}_{\text{ICI}}$
$h_{\mathbf{k}+3}$

Figure 3.1 : llustration of ICI rule for noisy k-means.

It is important to stress that the proposed method depends on a threshold term $C > 0$ in (3.32). This problem was studied in Spokoiny and Vial [2009] using the propagation method.

3.3.3 Anisotropic EGC method

The EGC (Empirical Gradient Comparison) rule is an anisotropic bandwidth selection rule. It was motivated in Section 3.2 where general adaptive properties have been stated in kernel empirical risk minimization problems. Here, we propose to use this method in clustering with errors-in-variables to illustrate the results of Section 3.2.

The EGC rule is based on the computation of gradient empirical risk as in (3.30) instead of empirical risk as in ERC. The principal motivation to use the gradient is summarized in Section 3.2, where EGC rule is described precisely. In the context of noisy clustering, the data-driven bandwidth is defined as :

$$\widehat{h}_{\text{EGC}} = \arg\min_{h \in \mathcal{H}_{\text{aniso}}} \widehat{\text{BV}}(h), \tag{3.34}$$

where $\widehat{\text{BV}}(h)$ is an estimation of the bias-variance decomposition of the excess risk. This quantity is based on the introduction of an auxiliary kernel :

$$\widetilde{\mathcal{K}}_{h,h'} = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}_h * \mathcal{K}_{h'}]}{\mathcal{F}[\eta]}\right](x),$$

where $\mathcal{K}_h * \mathcal{K}'_h$ stands for the convolution product between two kernel functions. This auxiliary kernel allows to compute the auxiliary gradient empirical risk $\widehat{G}_{h,h'}(\mathbf{c})$, where $\widetilde{\mathcal{K}}_{h,h'}$ is used in (3.30) instead of $\widetilde{\mathcal{K}}_h$. Then, the quantity $\widehat{\mathrm{BV}}(h)$ in (3.34) is defined as :

$$\widehat{\mathrm{BV}}(h) := \sup_{h'\in\mathcal{H}_{\mathrm{aniso}}} \left\{ |\widehat{G}_{h,h'} - \widehat{G}_{h'}|_{2,\infty} - \mathcal{M}_l(h,h') \right\} + \mathcal{M}_l^{\infty}(h), \quad \text{with } \mathcal{M}_l^{\infty}(h) := \sup_{h'\in\mathcal{H}_{\mathrm{aniso}}} \mathcal{M}_l(h',h),$$

where $|T|_{2,\infty} := \sup_\theta |T(\theta)|_2$ for any $T : \mathbb{R}^{dk} \to \mathbb{R}dk$ whereas $\mathcal{M}_l(h,h')$ is a majorant function according to Definition 7. In our framework, it is defined in Lemma 8 for the mildly ill-posed case as :

$$\mathcal{M}_l(h,h') = C_{\mathrm{aniso}}\sqrt{kd}\left(\frac{\Pi_{i=1}^d h_i^{-\beta_i}}{\sqrt{n}} + \frac{\Pi_{i=1}^d (h_i \vee h_i')^{-\beta_i}}{\sqrt{n}}\right), \tag{3.35}$$

where $C_{\mathrm{aniso}} > 0$ is a positive constant. Note that in this experimental study, we also consider a Gaussian distribution for the noise ϵ in (3.29). In this case, we choose a majorant function in $\widehat{\mathrm{BV}}(h)$ as a product of exponentially decreasing functions of h_i, $i = 1\dots,d$ instead of polynomial type as in Lemma 8. This choice is originated in Comte and Lacour [2013] where a study of the standard GL method is suggested in a deconvolution setting.

The computation of (3.34) requires many optimization steps. To overcome this computational issue, in our simulations we use simultaneously packages doParallel and foreach to provide a parallel execution of our R code on machines with multiples cores. The foreach package promotes a new looping construct for executing R code repeadtly. It is similar to the standard lapply function, but does not require the evaluation of a function. It facilitates the execution of the loop in parallel. The doParallel package registers the parallel backend with the foreach package. In our simulation study, we use a machine with 64 cores to speed up the EGC minimization (3.34).

3.3.4 Experiments

We consider the experimental setting of Section 2.3.2 with $j = 1$, that is we restrict to the case $k = 2$ for simplicity. We generate an i.i.d. noisy sample $\mathcal{D}_n = \{Z_1,\dots,Z_n\}$ where :

$$Z_i = X_i + \epsilon_i(u),\ i = 1,\dots,n, \tag{3.36}$$

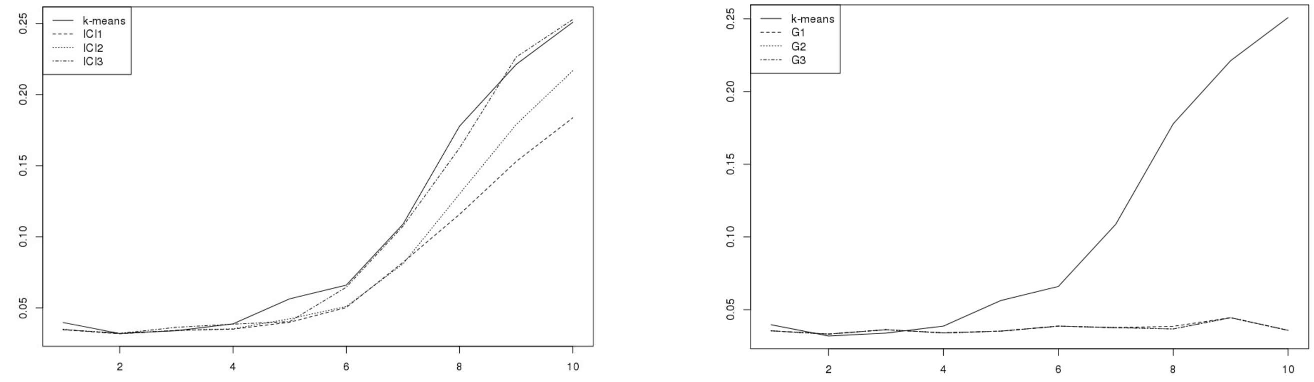
where $(X_i)_{i=1}^n$ are i.i.d. with density f defined as :

$$f^{(1)} = 1/2 f_{\mathcal{N}(0_2,I_2)} + 1/2 f_{\mathcal{N}((5,0)^T,I_2)}.$$

In this study, $(\epsilon_i(u))_{i=1}^n$ are i.i.d. with Gaussian noise with zero mean $(0,0)^T$ and covariance matrix $\Sigma(u) = \begin{pmatrix} 1 & 0 \\ 0 & u \end{pmatrix}$ for $u \in \{1,\dots,10\}$. In this setting, we propose to compare the performances of k-means with Noisy k-means by computing the clusterring error (see (2.43)) and the quantization error (see (2.44)).

For each criterion, we study the behaviour of the Lloyd algorithm (standard k-means) with two different noisy k-means, corresponding to two different choice of bandwidths h, with ERC of Section 3.3.2 or EGC of Section 3.3.3. Thanks to the theoretical results, we know that each bandwidth selection method depends on some constant. For ERC with ICI implementation, the constant $C_{\mathrm{iso}} > 0$ appears in (3.32) whereas for the gradient, $C_{\mathrm{aniso}} > 0$ is involved in (3.35). In the sequel, we illustrate the behaviour of these methods with respect to the fluctuation of these constants.

Clustering risk Figure 3.2 (a)-(b) illustrates the evolution of the clustering risk (2.43) when $u \in \{1,\dots,10\}$ in the model for k-means and the two selection rules. For each rule, we bring into play three different constants.

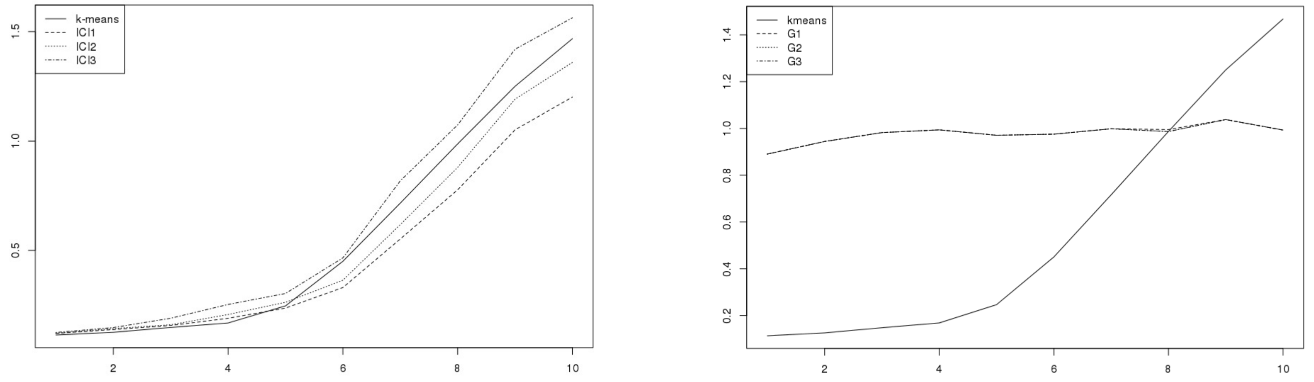


(a) ERC method (b) Gradient method

Figure 3.2 : Clustering risk averaged over 100 replications with $n = 200$ for k-means against ICI (a) and the gradient (b).

The performances of ERC method with ICI implementation depends strongly on the constant $C_{\text{iso}} > 0$ which appears in (3.32). A good calibration of this constant gives slightly better results than k-means (Figure 3.2 (a)). In comparison, the noisy k-means algorithm with EGC method significantly outperforms k-means or ERC (Figure 3.2 (b)). That highlights the importance in practice to choose two different bandwidths in each direction in this model, i.e. an anisotropic bandwidth. Moreover, the calibration of ERC is more difficult than EGC, which confirms the theoretical study of this chapter.

Quantization risk In Figure 3.3, we plot the quantization error of each procedure, when the variance of the noise increases. As before, we employ different constants for each method.



(a) ERC method (b) Gradient method

Figure 3.3 : Quantization risk averaged over 100 replications with $n = 200$.

Adaptive Noisy k-means with ICI do not show a good accuracy in terms of quantization error whereas EGC rule is better. It also shows one more time that quantization is harder than clustering.

3.3.5 Conclusion of the experimental study

This section investigates the bandwidth selection problem in noisy k-means. By using theoretical results of Section 3.1 and Section 3.2, we present two data-driven bandwidth selection for both the isotropic and anisotropic case. A first simulation study reveals a good behaviour of EGC in terms of clustering. Many other problems could be adressed in the future. One can use these bandwidth selection methods in other kernel empirical risk minimization problems, such as in image denoising or local fitted likelihood. In particular, it could be a way of calibrating a local constant approximation method in image denoising with non gaussian noise by using robust loss, such as the Huber loss.

From i.i.d. learning to online learning

Chapters 2-3 are an invitation to the statistical learning theory. By considering a contaminated sample, we had the opportunity to (1) rewrite the theory of empirical process of van de Geer [2000] (2) generalize the classical lower bounds of Mammen and Tsybakov [1999] and Audibert and Tsybakov [2007] (3) introduce an alternative to localization into the statement of fast rates of convergence. Inverse Statistical Learning generates a tedious problem of bandwidth selection that we solve with the inspiration of Lepski's contributions. To deal with the isotropic case, we compare empirical risks instead of estimators and give adaptive fast rates for the excess risk. In the anisotropic case, Goldenshluger and Lepski's principle (see Goldenshluger and Lepski [2011]) is applied to the comparison of gradient empirical risks. From the practical viewpoint, these considerations are illustrated in clustering. It gives one novel noisy clustering algorithm called noisy k-means, and two novel bandwidth selection methods, called ERC (Empirical Risk comparison) and EGC (Empirical Gradient Comparison).

As mentioned earlier in the dissertation, two paradigms are usually exposed in learning theory. In the two previous chapters, we suppose we have at our disposal an i.i.d. sample of random variables. It diriges the theoretical effort on uniform bounds for empirical processes, empirical risk minimizers, as well as excess risk bounds. However, another popular depiction of a statistical problem could be investigated. This is the purpose of Chapter 4, where online learning is studied. In this framework, the data arrives sequentially without any probabilistic assumptions. It gives us a new and appealing source of theoretical issues, in terms of algorithms, regret bounds, and adaptivity. This playground is investigated in the next pages for the problem of clustering, where a theoretical study in this case is unusual.

Chapitre 4

0nline learning with sparsity priors

We know that ℓ_0-penalized methods enjoy good theoretical properties but untractable computational complexity. On the contrary, convex relaxations, such as the lasso, reach sparsity oracle inequalities under rather restrictive assumptions.

To tackle this impasse, Dalalyan and Tsybakov [2012] come up with sparsity priors in a fairly general set-up. To set this idea, let us consider an i.i.d. sample $\mathcal{D}_n = \{(X_i, Y_i),\ i = 1, \dots, n\}$ with law P_f, where the regression function $f(\cdot) = \mathbb{E}(Y|X = \cdot) \in \mathcal{F}_\Theta = \{f_\theta, \theta \in \Theta\}$, for $\Theta \subseteq \mathbb{R}^p$ with p possibly larger than n. In this setting, PAC-Bayesian inequalities are affirmed on the following basis :

$$\mathbb{E}_{\mathcal{D}_n} \|\hat{f} - f\|^2_{L^2(P_X)} \leq \inf_{\rho \in \Delta(\Theta)} \left\{ \mathbb{E}_{\theta' \sim \rho} \|f_{\theta'} - f\|^2_{L^2(P_X)} + \frac{\mathcal{K}(\rho, \pi)}{\lambda(n+1)} \right\}, \tag{4.1}$$

where $\Delta(\Theta)$ is the set of probability distribution over Θ, $\mathcal{K}(\cdot,\cdot)$ is the Kullback-Leibler divergence and $\lambda > 0$ is an inverse temperature parameter. In the previous inequality, $\hat{f} := \mathbb{E}_{\theta \sim \hat{\rho}} f_\theta$ where $\hat{\rho} := \hat{\rho}(\mathcal{D}_n)$ is a random measure based on a mirror averaging :

$$\hat{\rho}(d\theta) = \frac{1}{n+1} \sum_{i=0}^{n} \hat{\rho}_i(d\theta) \text{ where } \hat{\rho}_i(d\theta) = \frac{e^{-\lambda \sum_{j=1}^{i} (Y_j - f_\theta(X_j))^2}}{\mathbb{E}_{\theta' \sim \pi} e^{-\lambda \sum_{j=1}^{i} (Y_j - f_{\theta'}(X_j))^2}} \pi(d\theta).$$

Inequality (4.1) holds for any choice of prior π. Then, in order to reach sparsity oracle inequalities, the following sparsity prior is introduced :

$$\pi(d\theta) = \prod_{j=1}^{p} a_\tau \left(\tau^2 + \theta_j^2\right)^{-2} d\theta_j, \tag{4.2}$$

where $a_\tau > 0$ is a normalizing constant. Sparsity priors have been introduced in Bayesian estimation by several authors (Johnstone and Silverman [2005], Rivoirard [2006], Seeger [2008]). The principle is to employ heavy-tailed distribution, such as multivariate Laplace, quasi-Cauchy or Pareto priors. In this PAC-Bayesian frawework, prior (4.2) guarantees sparsity oracle inequalities thanks to (4.1).

More recently, sparsity priors have been used in online learning. Given a deterministic sequence (x_t, y_t), $t = 1, \dots, T$ where $x_t \in \mathbb{R}^d$ and $y_t \in \mathbb{R}$, a dictionary of base forecasters $(\varphi_k)_{k=1}^p$ defined in $\mathbb{R}^d$, Gerchinovitz [2013] produces a sequential algorithm satisfying a sparsity regret bound :

$$\sum_{t=1}^{T} (\hat{y}_t, y_t)^2 \leq \inf_{\theta \in \mathbb{R}^p} \left\{ \sum_{t=1}^{T} (y_t - f_\theta(x_t))^2 + \frac{|\theta|_0}{\lambda} \log\left(1 + \frac{|\theta|_1}{|\theta|_0 \tau}\right) \right\} + \tau^2 B_\Phi, \tag{4.3}$$

where $B_\Phi > 0$, $|\cdot|_0$ is the ℓ_0-norm in $\mathbb{R}^p$ and $f_\theta(\cdot) = \sum_{k=1}^{p} \theta_k \varphi_k(\cdot)$. In (4.3), for each $t \geq 1$, $\hat{y}_t := \mathbb{E}_{\hat{\rho}_t} f_\theta(x_t)$ where :

$$\hat{\rho}_t(d\theta) = \frac{e^{-\lambda \sum_{j=1}^{t-1} (y_j - f_\theta(x_j))^2}}{\mathbb{E}_{\theta' \sim \pi} e^{-\lambda \sum_{j=1}^{t-1} (y_j - f_{\theta'}(x_j))^2}} \pi(d\theta), \tag{4.4}$$

with π a sparsity prior such as (4.2).

Prediction with expert's advices is the core of a huge amount of work these last decades in game theory and statistics (see Cesa-Bianchi and Lugosi [2006] and the references therein). The problem could be described as a sequential game between the nature and a forecaster. A blackbox - the nature - reveals at each trial t a real value $y_t \in \mathbb{R}$. Then, the forecaster predicts the next value based on the past observations and expert advices. These expert advices could be based on deterministic - or stochastic - models, or even adversarial. The goal is to predict as well as the best expert, no matter what sequence is produced by the blackbox. Very often, the introduced algorithms are based on convex combinations of expert advices, where coefficients depend on the past performances of each expert as in (4.4). In this respect, exponential weights are very often introduced. Applications of this online scenario are ubiquitous : they include weather forecasting, finance, social sciences, time series, etc...

In this chapter, we turn out into an unsupervised counterpart ot this problem, where we want to predict a multivariate individual sequence with no expert advices. In the context of clustering, we intent to relax the assumptions on the data-generating mechanism introduced in statistical learning (see Chapter 2 and Chapter 3). We give algorithms with theoretical guarantees for the problem of online clustering without any probabilistic hypothesis. The proposed algorithms are exponential weighting procedures inspired from Dalalyan and Tsybakov [2012] and Gerchinovitz [2013]. In the first section of this chapter, we state sparsity regret bounds for a fully automatic PAC-Bayesian sequential algorithm with sparsity prior derived from (4.2). This methodology also offers surprising new insights into the classical i.i.d. setting thanks to the now popular "online to batch conversion". We illustrate this technique in model selection clustering as well as high dimensional clustering. Then, we extend this approach to the problem of online bi-clustering. In this case, the problem of online clustering is a step into a high-level task of online prediction as in Gerchinovitz [2013]. We establish sparsity regret bounds where the sparsity is related with the structure of the data points. Eventually, we study the minimax regret for these problems and offer lower bounds under a sparsity scenario. Lower bounds are derived from a simple probabilistic reduction scheme as in Haussler, Kivinen, and Warmuth [1998]. Surprisingly, these bounds match - at least asymptotically - with upper bounds under the worst case scenario.

4.1 Online clustering of individual sequences [L9]

4.1.1 Introduction

In this section, we construct online *clustering* algorithms which learn according to the following protocol. On each day t, the forecaster must predict the next instance $x_t \in \mathbb{R}^d$ with at most $p \geq 1$ possible "proposals" or "strategies". On the morning of day t, he has access to the inputs $x_1, \dots, x_{t-1}$ of the previous days. Based on these instances, he must propose a codebook of $p \geq 1$ strategies $\hat{\mathbf{c}}_t = (\hat{c}_{t,1}, \dots, \hat{c}_{t,p}) \in \mathbb{R}^{dp}$. At the end of the day, he receives x_t and incurs a loss - or distortion - $\ell(\hat{\mathbf{c}}_t, x_t)$, where :

$$\ell(\hat{\mathbf{c}}_t, x_t) = \min_{j=1,\dots,p} |\hat{c}_{t,j} - x_t|_2^2,$$

and $|\cdot|_2$ stands for the Euclidean norm in $\mathbb{R}^d$. The goal of the forecaster is to control the cumulative distortion $\sum_{t=1}^T \ell(\hat{\mathbf{c}}_t, x_t)$, with $|\hat{\mathbf{c}}_t|_0$ as small as possible, where $|\hat{\mathbf{c}}_t|_0$ corresponds to the number of non-zero strategies at time t, i.e. :

$$|\mathbf{c}|_0 := \text{card}\{j = 1, \dots, p : c_j \neq (0, \dots, 0)^\top \in \mathbb{R}^d\}, \ \forall \mathbf{c} = (c_1, \dots, c_p) \in \mathbb{R}^{dp}. \tag{4.5}$$

Before all else, let us describe a tedious candidate strategy. At each trial $t \geq 1$, the forecaster deposits a proposal on each past instance $x_1, \dots, x_t \in \mathbb{R}^d$ and let the other components to zero. This tiresome system will satisfy $|\hat{\mathbf{c}}_t|_0 = t$, for any $t \geq 1$. Consequently, the strategy will induce small cumulative loss $\sum_{t=1}^T \ell(\hat{\mathbf{c}}_t, x_t)$ but huge complexity, and is equivalent to the so-called "overfitting phenomenon". In this contribution, we want to develop algorithms which summarize the information of the deterministic sequence, namely such that $|\hat{\mathbf{c}}_t|_0 << t$.

A motivating example[1] is as follows. A t-shirt sailer receives online data about their sales, customer after customer (such as prize, color and shape). After each checkout process t, he must predict the next instance in order to market appropriately to the current customer's patterns. Since different social clusters are involved (such as boys, teens, or gothics), he can advise different strategies or attempts. His own limitation is to come up with a finite - and as small as possible - number of strategies in order to summarize the demand (he has not access to an infinite store size). Additionally, since fashion changes over time, the retailer wants to learn in an online way.

In this chapter, we make no assumption about the data generating mechanism. Our results hold for a worst case sequence of instances. It allows to tackle non stationarity in the learning process and predict - or cluster - sequences of points with time-varying structure. This problem has been, as far as we know, very poorly treated in the literature. In the framework of expert advice, we can mention Choromanska and Monteleoni [2012], where different clustering algorithms are aggregated at each trial to get an online clustering algorithm. Zong [2005] investigates an online version of the spherical k-means algorithm. In the present dissertation we have not any expert advice (such as k-means). In particular, we have no idea about the number of clusters to use in the learning protocol.

From a theoretical viewpoint, we are interested in sparsity regret bounds introduced in (4.3). More precisely, we recommend to control the cumulative loss according to :

$$\sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{\mathbf{c}\in\mathbb{R}^{dp}} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \lambda|\mathbf{c}|_0 \right\} + r_\lambda(T), \tag{4.6}$$

where $|\cdot|_0$ is defined in (4.5), $r_\lambda(T)$ is a residual term and $\lambda > 0$ is a temperature parameter. It has to be calibrated in order to minimize the right hand side of (4.6). In other words, we want to control the regret of our sequential procedure to have not reached the compromise between fitting the data and compress the information (i.e. the infimum which appears in the right hand side). Going back to the t-shirt retailer example, it means that we are looking at a strategy that fits the customer's patterns as well as possible, but also which minimizes the number of offers. This compromise is of first interest in information theory and statistics.

Our algorithms are based on standard sequential randomized procedures introduced above (see also Audibert [2009]). In Dalalyan and Tsybakov [2012], it is noted that these methods are computationally feasible for relatively large dimensions of the problems, by using a so-called Langevin Monte-Carlo method. These computational aspects have been also considered in Alquier and Biau [2013] in a sparse single index model and in Alquier and Guedj [2013] in a sparse additive model. This direction is of first interest in clustering.

In this section, we present the sequential randomized algorithm and give the first sparsity regret bounds as in (4.6) for the problem of online clustering with known horizon T. The problem of adaptation, namely the knowledge of T, is also studied and we give a fully automatic online algorithm where the temperature parameter $\lambda > 0$ and the prior scale $\tau > 0$ are calibrated automatically. Eventually, thanks to the well-known online-to-batch conversion, we illustrate the power of the PAC-Bayesian theory in the standard i.i.d. case. Using a slightly modified sequential algorithm, we perform model selection clustering as well as high dimensional clustering.

4.1.2 The algorithm of online clustering

For any integer $d, p \geq 1$, we denote by $\Delta(\mathbb{R}^{dp})$ the set of probability measure on $\mathbb{R}^{dp}$. Let us introduce a prior $\pi \in \Delta(\mathbb{R}^{dp})$ and an inverse temperature parameter $\lambda > 0$. At the beginning of the game, we draw $\hat{\mathbf{c}}_1$ with law $\hat{p}_1 := \pi$. We fix $S_0 \equiv 0$. Then, learning proceeds as the following sequence of trials $t = 1, \ldots, T-1$:

1. Get x_t and compute $S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda}{2}[\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2,\ \forall \mathbf{c} \in \mathbb{R}^{dp}$.
2. Let $\hat{p}_{t+1}(d\mathbf{c}) := \frac{e^{-\lambda S_t(\mathbf{c})}}{W_t}\pi(d\mathbf{c}) \in \Delta(\mathbb{R}^{dp})$, where $W_t = \mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda S_t(\mathbf{c})}$.

[1] This toy example was found in the MIT Opencourseware. Recently, I visited a company and concluded a CIFRE with a very closely related problem of e-commerce website.

3. Draw $\hat{\mathbf{c}}_{t+1}$ according to the law $\hat{p}_{t+1}$.

W have constructed a vector of probability measures $(\hat{p}_1, \dots, \hat{p}_T)$, where each $\hat{p}_t \in \Delta(\mathbb{R}^{dp})$ is calculated thanks to the sequence of past instances $x_1, \dots, x_{t-1}$ and the realizations of $(\hat{\mathbf{c}}_1, \dots, \hat{\mathbf{c}}_{t-1})$. More precisely, the principle is to update the current error of any codebook $\mathbf{c} \in \mathbb{R}^{dp}$ as follows :

$$S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda}{2}[\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2, \tag{4.7}$$

where $\lambda > 0$ is some temperature parameter. At each trial t, the loss of a codebook is decomposed as the loss over the past, the current loss $\ell(\mathbf{c}, x_t)$ and a stability term that ensures $\hat{\mathbf{c}}_{t+1}$ to be not so far from $\hat{\mathbf{c}}_t$. This term can be viewed as a penalization term to better control the variance in our procedure (see Audibert [2009] for details and inequality (4.10) below). Due to the construction of a randomized estimator $\hat{\mathbf{c}}$, we are interested in the cumulative expected loss, given by :

$$\mathcal{E}_T(\hat{\mathbf{c}}) := \sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_t)} \ell(\hat{\mathbf{c}}_t, x_t), \tag{4.8}$$

where for each $t \geq 1$, the product measure $(\hat{p}_1, \dots, \hat{p}_t)$ is constructed in the algorithm.

PAC-Bayesian bounds go back to the work of Mac Allester [1998] (see also Catoni [2001] or more recently Seeger [2008]). It gives a control in expectation of the risk of any randomized estimator. The precise expression of the upper bounds depends on the context, but it is very often an empirical risk penalized in terms of Kullback-Leibler divergence. A nice property is the following *duality formula*. For any measurable function $h : \mathbb{R}^{dp} \to \mathbb{R}$, we have :

$$\log \mathbb{E}_{\mathbf{c} \sim \pi} e^{h(\mathbf{c})} = \sup_{\rho \in \Delta(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\mathbf{c} \sim \rho} h(\mathbf{c}) - \mathcal{K}(\rho, \pi) \right\}. \tag{4.9}$$

Since the earlier work of Mac Allester, many authors have investigated PAC-Bayesian bounds. For our purpose, we can mention Audibert [2009], which has largely inspired the result of Proposition 1 below. In particular, the construction of the algorithm - and more precisely the update of the current error described in (4.23) - warrants :

$$\forall \lambda > 0, \forall t \geq 1, \ \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq -\frac{1}{\lambda} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_t)} \log \mathbb{E}_{\mathbf{c} \sim \rho} e^{-\lambda(S_t(\mathbf{c}) - S_{t-1}(\mathbf{c}))}. \tag{4.10}$$

Assertion (4.10) emanates from Audibert [2009] in a quite general setting as a variance inequality. This kind of inequality can be traced back to Haussler, Kivinen, and Warmuth [1998] (see also Juditsky, Rigollet, and Tsybakov [2008] in the i.i.d. setting). In our framework, it is the starting point to get the following result :

Proposition 1. *For any deterministic sequence $(x_t)_{t=1}^T \in \mathbb{R}^{dT}$, for any $p \in \mathbb{N}^\star$, any $\lambda > 0$ and any prior $\pi \in \mathcal{M}_+(\mathbb{R}^{dp})$, the cumulative loss* (4.8) *satisfies :*

$$\mathcal{E}_T(\hat{\mathbf{c}}) \leq \inf_{\rho \in \Delta(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\boldsymbol{c} \sim \rho} \sum_{t=1}^{T} \ell(\boldsymbol{c}, x_t) + \frac{\mathcal{K}(\rho, \pi)}{\lambda} + \frac{\lambda}{2} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_T)} \mathbb{E}_{\boldsymbol{c} \sim \rho} \sum_{t=1}^{T} [\ell(\boldsymbol{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2 \right\}. \tag{4.11}$$

The bound of Proposition 1 gives a control of the expected cumulative loss of the randomized procedure. It holds for any choice of prior π, as well as any inverse temperature parameter $\lambda > 0$. In the sequel, we choose a group-sparsity prior to give a sparsity regret bound for our problem.

Proof: Applying (4.10) for any t and summing over t, we can prove easily :

$$\mathcal{E}_T(\hat{\mathbf{c}}) \leq -\frac{1}{\lambda} \sum_{t=1}^{T} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_t)} \log \mathbb{E}_{\mathbf{c} \sim \hat{p}_t} e^{-\lambda[S_t(\mathbf{c}) - S_{t-1}(\mathbf{c})]} =: \zeta_T$$

We are now on time to bring into play the chain rule (see Barron [1987]) to get :

$$\zeta_T = -\frac{1}{\lambda} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_T)} \log \prod_{t=1}^{T} \left(\frac{W_t}{W_{t-1}} \right) = -\frac{1}{\lambda} \mathbb{E}_{(\hat{p}_1, \dots, \hat{p}_T)} \log \left(W_T \right).$$

The Kullback duality formula (4.9) applied with $h(\cdot) = -\lambda S_T(\cdot)$ concludes the proof. ■

4.1.3 Sparsity regret bounds

Group-sparsity encourages occurences of whole blocks of zeros in a decision vector (see Yuan and Lin [2007]). It has been used in many applications, such as genetics or image annotation (see Zhang, Huang, Huang, Yu, Li, and Metaxas [2010]), where the lasso is not consistent for variable selection in high correlation settings. In this chapter, we are looking at a vector $\mathbf{c} = (c_1, \dots, c_p) \in \mathbb{R}^{dp}$ such that $|\mathbf{c}|_0 := \text{card}\{j = 1, \dots, p : c_j \neq (0, \dots, 0)^\top\}$ is small, namely a so-called group-sparsity. To deal with such a property, we introduce a new kind of prior. It consists of a product of multivariate Student's distribution $\sqrt{2}\tau\mathcal{T}_d(3)$, where $\tau > 0$ is a scaling parameter and $\mathcal{T}_d(3)$ is the d-multivariate Student with three degrees of freedom. It can be viewed as a generalization of the prior in (4.2) where a product of univariate Student is considered. Consequently, we use a multivariate Student's distribution defined in Kotz and Nadarajah [2004], defined as the ratio between a gaussian vector and the square root of an independent χ^2 distribution with 3 degrees of freedom. In our case, it leads to the following representation :

$$\pi_S(d\mathbf{c}) := \prod_{j=1}^{p} \left\{ a_{R,\tau}^{-1} \left(1 + \frac{|c_j|_2^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \mathbf{1}(|c_j|_2 \leq 2R) \right\} d\mathbf{c}, \tag{4.12}$$

where $a_{R,\tau} := b_{d,\tau}\mathbb{P}(\sqrt{2}\tau|\mathcal{T}_d(3)|_2 \leq 2R)$ for some constant $b_{d,\tau} > 0$. Here, $R > 0$ is a threshold that could be chosen arbitrarily big. Roughly speaking, the scaling parameter $\tau > 0$ - which can be fixed to a really small parameter - ensures sparsity for the vector of p groups $\sqrt{2}\tau\mathcal{T}_d(3)$ whereas the heavy tails property of $\mathcal{T}_d(3)$ guarantees that a small proportion of groups are quite far from zero. From a theoretical viewpoint, the introduction of the group-sparsity prior (4.12) gives rise to the following lemma :

Lemma 10. *Let $p \in \mathbb{N}^*$, $\tau, R > 0$ and π_S defined in* (4.12)*. Let $\boldsymbol{c} = (c_1, \dots, c_p) \in \mathcal{B}^p(R)$ where $\mathcal{B}^p(R) = \{\boldsymbol{c} = (c_1, \dots, c_p) : \forall j = 1, \dots, p, |c_j|_2 \leq R\}$. Introduce p_0 the following translated version of π_S with mean $\boldsymbol{c}$:*

$$p_0(d\boldsymbol{c}') = \prod_{j=1}^{p} \left\{ a_{R,\tau}'^{-1} \left(1 + \frac{|c_j' - c_j|_2^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \mathit{1}(|c_j' - c_j|_2 \leq R) \right\} d\boldsymbol{c}',$$

where here $a'_{R,\tau} := b_{d,\tau}\mathbb{P}(\sqrt{2}\tau|\mathcal{T}_d(3)|_2 \leq R)$. Then $p_0 << \pi_S$ and we have :

$$\mathcal{K}(p_0, \pi_S) \leq (3 + d)|\boldsymbol{c}|_0 \log\left(1 + \frac{\sum_{j=1}^{p}|c_j|_2}{\sqrt{6}\tau|\boldsymbol{c}|_0}\right) + \frac{12pd\tau^2}{R^2}.$$

The proof of the lemma is based on Dalalyan and Tsybakov [2012] and the properties of the multivariate Student's distribution (see Kotz and Nadarajah [2004]).

Theorem 15. *For any deterministic sequence $(x_t)_{t=1}^T$, any $R > 0$, let us consider the online algorithm with $\lambda = \sqrt{(3 + d)/T}$ using prior π_S defined in* (4.12) *with $\tau^2 \leq (1/p) \wedge (1/\sqrt{T})$. Then :*

$$\begin{aligned} \mathcal{E}_T(\hat{\boldsymbol{c}}) \leq \inf_{\boldsymbol{c} \in \mathcal{B}^p(R)} &\left\{ \sum_{t=1}^{T} \ell(\boldsymbol{c}, x_t) + |\boldsymbol{c}|_0 \sqrt{T(3 + d)} \log\left(1 + \frac{\sqrt{T}\sum_{j=1}^{p}|c_j|_2}{\sqrt{6}|\boldsymbol{c}|_0}\right)\right\} \\ &+ \sqrt{T}\left(\frac{12d}{R^2\sqrt{3 + d}} + 8C^2R^2\sqrt{3 + d} + 6d\right). \end{aligned}$$

The choice of (τ, λ) above gives rise to a sparsity regret bound with rate $\mathcal{O}(\sqrt{T}\log T)$. The RHS of this regret bound does not depend on p, provided that $\tau > 0$ is chosen adequately. If we suppose the existence of a minimizer $\mathbf{c}^\star$ of the RHS of Theorem 15 such that $|\mathbf{c}^\star|_0 = s$ for some sparsity index $s \in \mathbb{N}^*$, we have, for T large enough :

$$\mathcal{E}_T(\hat{\mathbf{c}}) - \mathcal{E}_T(\mathbf{c}^\star) \leq \text{const.} \times s\sqrt{T}\log T.$$

Section 4.3 studies the optimality of this bound in a minimax sense.

From the adaptive point of view, the choice of the couple (λ, τ) depends explicitly on the horizon T, which is not known in a pure online setting. This problem is considered in Section 4.1.4 where a fully automatic version of the algorithm of Section 4.1.2 is given.

A tuning of R, as well as a more involved choice of $\lambda, \tau > 0$ are conceivable. It is likely to let $\lambda, \tau > 0$ depend on $B_T = \max_{t=1,\dots,T} |x_t|_2$ and $R > 0$ to get better constants in Theorem 15. More precisely, for $\lambda = \sqrt{(3+d)/(8T)}(2B_T + 3R)^{-1}$, we get a better residual term in the sparsity regret bound. However, this choice of λ entails a more difficult automatic choice since as in Gerchinovitz [2013], the parameter B_T has to be calibrated at each iteration in this case.

Eventually, the infimum in Theorem 15 is restricted to $\{\mathbf{c} : \forall j,\ |c_j|_2 \leq R\}$. This arises for technicalities in the proof and could be extended to the whole space $\mathbb{R}^{dp}$ in the spirit of Gerchinovitz [2013].

4.1.4 Adaptation

The choice of the tuning parameters $\lambda, \tau > 0$ in Theorem 15 depends explicitly on the horizon T of the deterministic sequence. However, if we consider a pure online setting, the size of the deterministic sequence is unknown. This problem is called adaptation in the deterministic literature and has been extensively studied in the context of prediction with expert advices (see Auer, Cesa-Bianci, and Gentile [2002], Cesa-Bianchi, Mansour, and Stoltz [2007], Gerchinovitz [2013]). Originally, one can use a doubling trick, which consists in restarting the algorithm at periods of exponentially increasing lengths of size 2^k, for $k \geq 1$. A more natural alternative is to let the tuning parameters depend on the trial $t \geq 1$. The idea has been introduced in Auer, Cesa-Bianci, and Gentile [2002] and influences the regret bounds by only a constant factor. These approaches are developed below for tuning both parameters λ and τ in the algorithm.

Adaptive temperature algorithm

Let us consider a prior $\pi \in \Delta(\mathbb{R}^{dp})$ and a non-increasing sequence of positive temperature parameter $(\lambda_t)_{t=1}^T$. At the beginning of the game, we draw $\hat{\mathbf{c}}_1$ with law $\hat{p}_1 := \pi$. We fix $S_0 \equiv 0$. Then, learning proceeds as the following sequence of trials $t \in \{1, \dots, T-1\}$:

1. Get x_t and compute : $S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda_t}{2}[\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2,\ \forall \mathbf{c} \in \mathbb{R}^{dp}$.
2. Let $\hat{p}_{t+1}(d\mathbf{c}) := \frac{e^{-\lambda_{t+1} S_t(\mathbf{c})}}{W_t}\pi(d\mathbf{c})$ where $W_t = \mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda_{t+1} S_t(\mathbf{c})}$.
3. Draw $\hat{\mathbf{c}}_{t+1}$ according to the law $\hat{p}_{t+1}$.

The submitted algorithm is denoted as $\hat{\mathbf{c}}_\tau$. It depends on a sequence of non-increasing temperature parameters $(\lambda_t)_{t=1}^T$. By choosing $\lambda_t = \sqrt{(3+d)/t}$, we arrive at the following adaptive regret bound.

Theorem 16. *For any deterministic sequence $(x_t)_{t=1}^T$, any $\tau, R > 0$, let us consider the adaptive algorithm $\hat{\mathbf{c}}_\tau$ with $\lambda_t = \sqrt{(3+d)/t}$ and prior π_S defined in* (4.12)*. Then :*

$$\mathcal{E}_T(\hat{\boldsymbol{c}}_\tau) \leq \inf_{\boldsymbol{c} \in \mathcal{B}^p(R)} \left\{ \sum_{t=1}^T \ell(\boldsymbol{c}, x_t) + \sqrt{3+d}|\boldsymbol{c}|_0 \sqrt{T} \log\left(1 + \frac{\sum_{j=1}^p |\boldsymbol{c}_j|_2}{\sqrt{6}|\boldsymbol{c}|_0 \tau}\right)\right\}$$
$$+ 16C^2R^2\sqrt{T(3+d)} + \frac{12pd\tau^2}{R^2\sqrt{3+d}}\sqrt{T} + 6d\tau^2 T.$$

This result gives a sparsity regret bound when λ varies over time in the sequential procedure. Moreover, if we choose $p \geq \sqrt{T}$ and a scale parameter $\tau \leq p^{-1/2}$, we arrive at :

$$\mathcal{E}_T(\hat{\mathbf{c}}_\tau) \leq \inf_{\mathbf{c} \in \mathcal{B}^p(R)} \left\{ \sum_{t=1}^T \ell(\mathbf{c}, x_t) + (3+d)|\mathbf{c}|_0 \sqrt{T} \log\left(1 + \frac{\sqrt{p}\sum_{j=1}^p |\mathbf{c}_j|_2}{\sqrt{6}|\mathbf{c}|_0}\right)\right\}$$
$$+ \sqrt{T}\left(16C^2R^2 + \frac{12d}{R^2\sqrt{3+d}} + 6d\right). \tag{4.13}$$

However, these choices of (p, τ) are not possible in the adaptive setting of unknown horizon T. Note that this problem also occurs in Gerchinovitz [2013], where a doubling trick is performed to get a fully automatic algorithm. Last paragraph follows the same lines and suggests a fully automatic procedure of online clustering.

A fully automatic online clustering algorithm

The idea is to decompose the sequence of outcomes $(x_t)_{t\geq 1}$ into sequences of exponentially increasing length as follows. Let $t_0 = 0$ and introduce, for any $r \in \mathbb{N}^*$, an integer t_r defined as :

$$t_r = \min\{t \geq t_{r-1} + 1 : \log(1+\sqrt{t}) > 2^r\}.$$

Let $\tau(r) = (e^{2^r} - 1)^{-1}$. Then, we construct the sequence of posterior distributions as follows : for any $r \in \mathbb{N}^*$, build from the sequence $(x_{t_{r-1}}, \ldots, x_{t_r - 1})$ the vector of posterior $(\hat{p}_{r,t_{r-1}}, \ldots \hat{p}_{r,t_r-1})$ with the adaptive temperature algorithm with $\tau := \tau(r)$ and $p = t_r - t_{r-1}$. The associated sequence of randomized estimators is denoted as $(\hat{\mathbf{c}}_{t,*})_{t\geq 1}$, where :

$$\forall r = 1, \ldots \; \forall t \in \{t_{r-1}+1, \ldots, t_r\},\; \hat{\mathbf{c}}_{t,*} \sim \hat{p}_{r,t}.$$

For any $T \geq 1$, for any sequence $(x_t)_{t=1}^T$, we hence have constructed a randomized sequence $\hat{\mathbf{c}}_* = (\hat{\mathbf{c}}_{t,*})_{t=1}^T$ which does not depend on horizon T.

Theorem 17. *For any $T \geq 1$, for any deterministic sequence $(x_t)_{t=1}^T$, any $R > 0$, let us consider the adaptive algorithm $\hat{\mathbf{c}}_* = (\hat{\mathbf{c}}_{t,*})_{t=1}^T$ defined above. Then :*

$$\begin{aligned}\mathcal{E}_T(\hat{\mathbf{c}}_*) \leq \inf_{c \in \mathcal{B}^p(R)} &\left\{ \sum_{t=1}^T \ell(\boldsymbol{c}, x_t) + |\boldsymbol{c}|_0 \sqrt{T(3+d)} \log\left(1 + \frac{\sum_{j=1}^p |\boldsymbol{c}_j|_2}{\sqrt{6}|\boldsymbol{c}|_0 \tau}\right)\right\} \\ &+ c\sqrt{T} \log\log(1+\sqrt{T}).\end{aligned}$$

Proof: The poof is based on the decomposition of the cumulative expected distortion in each period $r \in \{1, \ldots, R\}$ as follows :

$$\sum_{t=1}^T \mathbb{E}_{(\hat{p}_1,\ldots,\hat{p}_t)} \ell(\hat{\mathbf{c}}_t, x_t) = \sum_{r=1}^R \sum_{t=t_{r-1}+1}^{t_r} \mathbb{E}_{(\hat{p}_{t_{r-1}+1},\ldots,\hat{p}_t)} \ell(\hat{\mathbf{c}}_t, x_t), \tag{4.14}$$

where $t_R = \min\{t \in (t_r)_{r\geq 1} : t > T\}$. Then we can apply Theorem 16 with $\tau = \tau(r)$, $p = t_r - t_{r-1}$ to get the result. ∎

4.1.5 Batch revisited

In this section, we go back to the vanilla clustering of an i.i.d. sample. Given an integer $k \geq 1$ and a probability P over $\mathbb{R}^d$, we write here :

$$\mathcal{W}_k(\mathbf{c}) = \mathbb{E}_P \min_{j=1,\ldots,k} |X - c_j|_2^2,\ \forall \mathbf{c} \in \mathbb{R}^{dk}.$$

In this setting, based on an i.i.d. sample $X_1, \ldots, X_n$, we introduce :

$$\widehat{\mathcal{W}}_k(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,k} |X_i - c_j|_2^2,\ \forall \mathbf{c} \in \mathbb{R}^{dk}. \tag{4.15}$$

Standard algorithms, such as the Lloyd algorithm, are made of Newton's type iterations and depend strongly on the initialization step. Moreover, the knowledge of k is not always guaranteed and a data-driven choice of this parameter remains a hard issue. In this paragraph, we suggest to use the PAC-Bayesian framework to get a fully automatic algorithm that performs model selection clustering. Eventually, for completeness, we also deal with the problem of high dimensional clustering. In this case, the number of clusters k is known but the dimension d of the variable X could be much larger than the sample size n.

Model selection clustering

Recently, Fischer [2011] formulates the problem of selecting the number of clusters k as a problem of model selection. She gives standard-style statistical learning bounds by using empirical process theory. For any integer $k \geq 1$, let us denote $\hat{\mathbf{c}}_k$ the minimizer of (4.15). Given the family $\{\hat{\mathbf{c}}_k, k = 1, \ldots, n\}$, Fischer [2011] suggests a penalized model selection procedure to choose k as follows :

$$\hat{k} = \arg \min_{k=1,\ldots,n} \left\{ \widehat{\mathcal{W}}_k(\hat{\mathbf{c}}_k) + \text{pen}_d(k) \right\},$$

where $\text{pen}_d(k)$ is an increasing function of the dimension kd. In practice, the choice of the penalty is made of two steps :

1. A theoretical study gives the shape of the penalty, namely here (see Fischer [2011]) :

$$\text{pen}_d(k) = \square \sqrt{\frac{kd}{n}}, \text{ for some } \square > 0.$$

2. Then, the constant $\square > 0$ in front of the penalty's shape can be calibrated thanks to the slope heuristic (see Baudry, Maugis, and Michel [2012]).

In this paragraph, we develop the PAC-Bayesian analysis for this problem. Let us introduce an integer $p \geq 1$, which could be large enough (we can choose $p = n$ to fix the ideas). Consider the prior $\pi_S \in \Delta(\mathbb{R}^{dp})$ defined according to (4.12) :

$$\pi_S(d\mathbf{c}) := \prod_{j=1}^{p} \left\{ a_{R,\tau}^{-1} \left(1 + \frac{|c_j|_2^2}{6\tau^2} \right)^{-\frac{3+d}{2}} \mathbf{1}(|c_j|_2 \leq 2R) \right\} d\mathbf{c}.$$

Fix $S_0 \equiv 0$ and draw $\hat{\mathbf{c}}_1$ according to π. Then, for any $i \in \{1, \ldots, n\}$:

1. Get X_i and compute : $S_i(\mathbf{c}) = S_{i-1}(\mathbf{c}) + \ell(\mathbf{c}, X_i) + \frac{\lambda}{2}[\ell(\mathbf{c}, X_i) - \ell(\hat{\mathbf{c}}_i, X_i)]^2, \ \forall \mathbf{c} \in \mathbb{R}^{dp}$.
2. Let $\hat{p}_{i+1}(dc) := \frac{e^{-\lambda S_i(\mathbf{c})}}{W_i} \pi(d\mathbf{c}) \in \Delta(\mathbb{R}^{dp})$ where $W_i = \mathbb{E}_{\mathbf{c}\sim\pi} e^{-\lambda S_i(\mathbf{c})}$.
3. Draw $\hat{\mathbf{c}}_{i+1}$ according to $\hat{p}_{i+1}$.

The final estimator in the i.i.d. case, denoted as $\hat{\mathbf{c}}_{\text{MA}}$, is a realization of the uniform law over $\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{n+1}\}$:

$$\hat{\mathbf{c}}_{\text{MA}} \sim \hat{\mu} := \mathcal{U}(\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{n+1}\}), \tag{4.16}$$

where $\hat{\mu}$ is the uniform law over the set of estimators $\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{n+1}\}$, conditionally to the training set $\mathcal{D}_n$. This additional step is called Mirror Averaging (MA) and has been used in the i.i.d. setting by many authors (see for instance Juditsky, Rigollet, and Tsybakov [2008], Dalalyan and Tsybakov [2012], Audibert [2009]). Since $\hat{\mathbf{c}}_{\text{MA}}$ is a realization of an uniform law, we are finally interested in the expectation (with respect to the training set $\mathcal{D}_n$) of the expected risk of $\hat{\mathbf{c}}_{\text{MA}}$, given by :

$$\mathbb{E}_{\mathcal{D}_n} \mathbb{E}_{\mathbf{c}'\sim\hat{\mu}} \mathcal{W}(\mathbf{c}') = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}_{\mathcal{D}_i} \mathbb{E}_{(\hat{p}_1,\ldots,\hat{p}_i)} \mathcal{W}(\hat{\mathbf{c}}_i) \text{ where } \mathcal{W}(\mathbf{c}) = \mathbb{E}_P \min_{c\in\mathbf{c}} |X - c|_2^2.$$

Theorem 18. *Suppose the distribution P satisfies $P(|X|_2 \leq B) = 1$ for some $B > 0$. Let us consider the mirror averaging $\hat{\mathbf{c}}_{\text{MA}}$ defined in (4.16) using parameters $R > 0$, $p = n$ and prior π_S defined in (4.12). If we choose $(\tau, \lambda) = \left(n^{-1/2}, n^{-1/2}\right)$, the following holds :*

$$\mathbb{E}_{\mathcal{D}_n} E_{\boldsymbol{c}'\sim\hat{\mu}} \mathcal{W}(\boldsymbol{c}') \leq \inf_{1\leq k\leq n} \left\{ \mathcal{W}(\boldsymbol{c}_k^\star) + \frac{(3+d)k}{\sqrt{n}} \log\left(1 + \frac{\sqrt{n}\sum_{j=1}^{n} |\boldsymbol{c}_j|_2}{\sqrt{6}k} \right) \right\}$$
$$+ n^{-1/2} \left(\frac{12d}{R^2} + 8C^2R^2 \right) + \frac{6d}{n},$$

where $\boldsymbol{c}_k^\star = \arg\min_{\boldsymbol{c}\in\mathcal{B}^p(R):|\boldsymbol{c}|_0=k} \mathcal{W}(\boldsymbol{c})$.

The RHS of Theorem 18 can be compared with Fischer [2011], where the penalized model selection procedure described above is used. The inequality of Theorem 18 makes sure that in the i.i.d. case, the risk of our procedure is comparable to the best codebook in the family, up to a residual term. This term approaches the rate $n^{-1/2}$, up to a $\log n$ factor. From a model selection point of view, if we compare this result with Fischer [2011], the main advantage of our approach is that there is not any tuning parameter to choose.

High dimensional clustering

In this paragraph, we turn out into the problem of high dimensional clustering (see Bouveyron and Brunet [2013], Parsons, Haque, and Liu [2004]). Given an integer $k \geq 1$, we consider an i.i.d. sample $X_1, \ldots, X_n$ with unknown law P over $\mathbb{R}^d$, where d could be much larger than n. We are interested in a codebook $\mathbf{c} = (c_1, \ldots, c_k)$ such that $|c_j|_0 << d$ for any $j = 1, \ldots, k$, where here, $|\cdot|_0$ stands for the usual ℓ_0-norm (i.e. the number of non-zero components in c_j). The main result of this paragraph is a sparsity oracle inequality for the mirror averaging estimator defined in (4.16) with a slightly different prior. In this setting of high dimensional clustering, we introduce the following sparsity prior :

$$\pi'_S(d\mathbf{c}) := \prod_{i=1}^{d} \left\{ a_{R,\tau}^{-1} \left(1 + \frac{|c_i|_2^2}{6\tau^2}\right)^{-\frac{3+k}{2}} \mathbf{1}(|c_i|_2 \leq 2R) \right\} d\mathbf{c}, \tag{4.17}$$

where $c_i = (c_{i1}, \ldots, c_{ik}) \in \mathbb{R}^k$ denotes the vector of the i^{th} coordinates of each c_j in $\mathbf{c} = (c_1, \ldots, c_k)$, and $a_{R,\tau} := b_{k,\tau}\mathbb{P}(\sqrt{2}\tau|\mathcal{T}_k(3)|_2 \leq 2R)$ for some constant $b_{k,\tau} > 0$. Let us briefly explain the introduction of this modified prior. Since we are looking at sparsity with respect to the dimension of the problem, we construct a product of d multivariate $\mathcal{T}_k(3)$ Student's distribution, where $k \geq 1$ is the known number of clusters in the problem. This choice mimics the introduction of π_S above. It encourages codebook $\mathbf{c}$ with small sparsity index $|\mathbf{c}|'_0$ defined as

$$|\mathbf{c}|'_0 = \text{card}\{i = 1, \ldots, d : (c_{i1}, \ldots, c_{ik}) \neq 0_{\mathbb{R}^k}\}.$$

Theorem 19. *Suppose distribution P satisfies $P(|X|_2 \leq B) = 1$. For some integer $k \geq 1$, let us consider the mirror averaging $\hat{\mathbf{c}}_{\text{MA}}$ defined in (4.16) using prior π'_S defined in (4.17) with parameters $R > 0$. If we choose : $(\tau, \lambda) = \left(d^{-1/2}, n^{-1/2}\right)$, the following holds :*

$$\mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{\boldsymbol{c}'\sim\hat{\mu}}R(\boldsymbol{c}') \leq \inf_{\boldsymbol{c}\in\mathcal{B}^d(R)} \left\{ \mathcal{W}(\boldsymbol{c}) + \frac{(3+k)|\boldsymbol{c}|'_0}{\sqrt{n}} \log\left(1 + \frac{\sqrt{d}\sum_{i=1}^d |\boldsymbol{c}_i|_2}{|\boldsymbol{c}|'_0}\right)\right\}$$
$$+ n^{-1/2}\left(\frac{12k}{R^2} + 8C^2R^2\right) + \frac{6k}{d}.$$

where $|\boldsymbol{c}|'_0 = card\{i = 1, \ldots, d : (c_{i1}, \ldots, c_{ik}) \neq 0_{\mathbb{R}^k}\}$ is the sparsity index of the codebook $\boldsymbol{c}$.

The RHS of Theorem 19 gives a rate of convergence of the form $\log d/\sqrt{n}$. The presence of a non-convex loss function gives rise to a rate of order $\mathcal{O}(n^{-1/2})$, up to a classical $\log d$ term.

4.2 Online bi-clustering with sparsity priors [L11]

Supervised classification is widely used for solving many real-world problems such as spam filtering or medical diagonostic (see Chapter 5). The main reason is the following : it can be easily expressed as a well-defined problem with a loss function. This loss function, chosen by the scientist, makes the problem clearly defined. On the other side, clustering is "unsupervised" classification, that is to assign classes that are not defined a priori. The goal is to learn the underlying structure of a dataset. This unsupervised problem remains a hard issue since this representation is not necessarily unique. Even worst, a huge number of different structures may coexist in any non-trivial set of data. In von Luxburg, Williamson, and Guyon [2009], two distinct purposes for clustering are expressed : data pre-processing where clustering is considered as a step in a whole data processing chain and exploratory data analysis to discover a new structure in a dataset. Earlier in the dissertation, exploratory data clustering was initiated in statistical and online learning. Clustering was carried out with the k-means loss function as a quantization problem : how to summarize a distribution P in terms of Euclidean distance ? In this section, a contrario, data pre-processing clustering is studied where the clustering task is an abstraction of the ultimate end-use problem.

4.2.1 Introduction

Bi-clustering or co-clustering is a popular method to analyse huge data matrices and build recommender systems. In this field, we mainly observe a random matrix, where rows correspond to a population and columns to variables (or products). This matrix is usually sparse, i.e. with many hidden entries (such as ratings). The goal is to reconstruct the matrix by clustering simultaneously the rows and colums of the matrix. This scenario has been applied to many real-world problems such as text mining (see Slonim and Tishby [2000]), gene expression (Cheng and Church [2000]), social networks (see Gnatyshak, Ignatov, Semenov, and Poelmans [2012]) or collaborative filtering (see Seldin [2009]). In Seldin and Tishby [2010], generalization bounds in terms of Kullback-Leibler divergence are proposed for this problem. Assuming the existence of a probabilistic distribution $p(x_1, x_2, y)$ over the triplet of rows, colums and rating, a discriminative predictor $q(y|x_1, x_2)$ is constructed via a PAC-Bayesian approach. The matrix is supposed to have i.i.d. entries and the number of clusters is known in advance. In this section, we want to investigate a deterministic and sequential version of the bi-clustering problem. As before, we are interested in sparsity regret bounds, where the sparsity is associated with the structure of the data points.

We consider an individual sequence (x_t, y_t), $t = 1, \ldots, T$ where T is the known horizon whereas for any $t = 1, \ldots, T$:

— the input variable $x_t = (x_{t,1}, \ldots, x_{t,d}) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_d =: \mathcal{X}$,

— the output [2] $y_t \in \mathcal{Y} \subseteq [0, M]$.

A seminal example is the construction of recommender systems. In this case, $d = 2$ and $x_t = (x_{t,1}, x_{t,2})$ corresponds to a couple customer×movie whereas y_t is the associated rating (such as $\{\star, \star\star, \star\star\star\}$ for instance). Note also that our analysis is not limited to the bi-clustering problem where $d = 2$ above, since we can consider $d > 2$ tensors as well. At each time t, an input $x_t \in \mathcal{X}$ is observed and we design a prediction $\hat{y}_t$. Then, y_t is given and we loss $\ell(\hat{y}_t, y_t) = (y_t - \hat{y}_t)^2$ for simplicity. This particular loss enjoys the useful property to be λ-exp-concave, which means that $\hat{y} \mapsto e^{-\lambda \ell(\hat{y}, y)}$ is concave. Unlike the previous pure unsupervised study, it allows to reach fast regret bounds (see also the discussion in Chapter 1). With this loss function, we mix decision functions according to :

$$\hat{y}_t = \mathbb{E}_{\vec{\mathbf{c}} \sim \hat{p}_t} g_{\vec{\mathbf{c}}}(x_t). \tag{4.18}$$

In (4.18), decision functions $g_{\vec{\mathbf{c}}}(\cdot)$ depend on a set of d-tensor codebooks $\vec{\mathbf{c}} = (\mathbf{c}_1, \ldots, \mathbf{c}_d) \in \prod_{j=1}^d \mathcal{X}_j^{p_j}$. A d-tensor codebook assigns to each component x_j of $x \in \mathcal{X}$ the nearest center of $\mathbf{c}_j := (c_{j,1}, \ldots, c_{j,p_j})$ for a given $p_j \in \mathbb{N}^*$. The associated d-tensor Voronoï cell is denoted as $V_{\vec{\mathbf{c}}}(x)$ and corresponds to the product of each Voronoï cell $V_{\mathbf{c}_j}(x_j) = \{z_j \in \mathcal{X}_j : \arg\min_{i_j=1,\ldots,p_j} |z_j - c_{j,i_j}|_2 = \arg\min_{i_j=1,\ldots,p_j} |x_j - c_{j,i_j}|_2\}$. In what follows (see for instance Theorem 20), we consider two different mappings $\vec{\mathbf{c}} \mapsto g_{\vec{\mathbf{c}}}(\cdot)$. The first one consists in computing the mean value of the sequence of past outputs in a given Voronoï cell. In this case, $g_{\vec{\mathbf{c}}}(x_t)$ is literally defined for any t as :

$$g_{\vec{\mathbf{c}}}^{\text{mean}}(x_t) = \frac{\sum_{u=1}^{t-1} y_u \mathbf{1}_{x_u \in V_{\vec{\mathbf{c}}}(x_t)}}{\text{card}\,\{\{x_1, \ldots, x_{t-1}\} \cap V_{\vec{\mathbf{c}}}(x_t)\}}. \tag{4.19}$$

In (4.19), we can initialize $g_{\vec{\mathbf{c}}}^{\text{mean}}(x_1) = M/2$ without loss of generality. In (4.19) (and also in (4.20)), $g_{\vec{\mathbf{c}}}(x_t)$ depends on the past observations $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$. We omit this dependence for simplicity. Furthermore, when $\mathcal{Y} = \{1, \ldots, M\}$, we can also use a majority vote for $g_{\vec{\mathbf{c}}}(x_t)$, where the majority vote at time t is taken in the Voronoï cell $V_{\vec{\mathbf{c}}}(x_t)$ as follows :

$$g_{\vec{\mathbf{c}}}^{\text{vote}}(x_t) = \arg\max_{k \in \mathcal{Y}} \text{card}\{u = 1, \ldots, t-1 : y_u = k \text{ and } x_u \in V_{\vec{\mathbf{c}}}(x_t)\}. \tag{4.20}$$

Equipped with these decision functions, we want to promote a sparse representation. Here, the sparsity is associated with the set of d-tensor codebooks. Given some vector of integers $\mathbf{m} = (m_1, \ldots, m_p) \in \mathbb{N}^p$, we restrict the study to the Euclidean space by considering $\mathcal{X}_j = \mathbb{R}^{m_j}$ for any $j = 1, \ldots, p$, $\mathcal{X} =$

[2] In the sequel, two cases are considered : $\mathcal{Y} = [0, M]$ (online regression) and $\mathcal{Y} = \{1, \ldots, M\}$ (online classification).

$\mathbb{R}^{\sum_{j=1}^d m_j p_j}$ and $\vec{\mathbf{c}} = (\mathbf{c}_1, \ldots, \mathbf{c}_d) \in \prod_{j=1}^d \mathbb{R}^{m_j p_j}$. Then, we wish that $y_t \approx g_{\vec{\mathbf{c}}^\star}(x_t)$, where $\vec{\mathbf{c}}^\star = (\mathbf{c}_1^\star, \ldots, \mathbf{c}_d^\star)$ is such that $\mathbf{c}_j^\star$ has a small ℓ_0-norm for any $j = 1, \ldots, d$ where $|\mathbf{c}_j^\star|_0$ is defined as :

$$|\mathbf{c}_j^\star|_0 = \text{card}\{i_j = 1, \ldots, p_j : c_{j,i_j}^\star \neq (0, \ldots, 0)^\top \in \mathbb{R}^{m_j}\}. \tag{4.21}$$

Consequently, we are looking for d distincts group-sparse codebooks $\mathbf{c}_1, \ldots, \mathbf{c}_d$. For this purpose, we will use in our algorithm a product of d group-sparsity priors introduced in Lemma 10 in online clustering. This prior is defined in Lemma 11 below.

As mentioned in the introduction, a comparable approach could be seen in Seldin and Tishby [2010] in a classical statistical learning context. By considering a random generator P with unknown probability distribution on the set $\mathcal{X} \times \mathcal{Y}$, Seldin and Tishby [2010] suggest to consider the following form of discriminative predictors :

$$h(y|x_1, \ldots, x_d) = \sum_{(i_1, \ldots, i_d)} h(y|i_1, \ldots, i_d) \Pi_{j=1}^d h(i_j|x_j).$$

In this stochastic setting, the hidden variables $(i_1, \ldots, i_d)$ represent the clustering of the input $X = (X_1, \ldots, X_d)$. Using a PAC-Bayesian analysis and bounds as in Mac Allester [1998], generalization errors in terms of Kullback-Leibler divergence are proposed. The randomized strategy is based on a density estimation of the law P, where the number of groups for each i_j, $j = 1, \ldots, d$ is known.

In this section, the framework is essentially different since we propose to use the results of Section 4.1 in order to get sparsity regret bounds according to :

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\vec{\mathbf{c}} \in \Pi_{j=1}^d \mathbb{R}^{m_j p_j}} \left\{ \sum_{t=1}^T (y_t - g_{\vec{\mathbf{c}}}(x_t))^2 + \text{pen}_0(\vec{\mathbf{c}}) \right\}, \tag{4.22}$$

where $\text{pen}_0(\vec{\mathbf{c}})$ is a penalty function which is proportional to the sum of the ℓ_0-norm (4.21) of the codebooks $\mathbf{c}_1, \ldots, \mathbf{c}_d$. The infimum in the RHS of (4.22) could be seen as a compromise between fitting the data and structural complexity, where the complexity is related to the number of clusters in each subspace $\mathcal{X}_j$, $j = 1, \ldots, d$. In recommender systems, it means that we can propose a simple representation of ratings with a block matrix with a few number of blocks. Note that this fact is strongly related with the usual sparsity assumption in low rank matrix completion (see for instance Candès and Recht [2009], Koltchinskii, Lounici, and Tsybakov [2011]).

4.2.2 General algorithm and associated PAC-Bayesian inequality

Before to describe the algorithm, let us introduce some notations. We denote by $\mathcal{C} := \Pi_{j=1}^d \mathbb{R}^{m_j p_j}$ the space of d-tensor codebooks, whereas a decision function at time t is denoted as $g_{\vec{\mathbf{c}}}(\cdot)$ (see (4.19) or (4.20) for instance). We introduce a prior $\pi \in \Delta(\mathcal{C})$, where $\Delta(\mathcal{C})$ is the set of probability measure on $\mathcal{C}$, and a temperature parameter $\lambda > 0$. We can now describe the general algorithm and its associated PAC-Bayesian inequality.

The principle of the algorithm is to predict y_t according to a mixture of decision functions $g_{\vec{\mathbf{c}}}$, where the mixture is updated by giving the best prediction of y_t at each iteration. At the beginning of the game, $\hat{p}_1 := \pi$. We observe x_1 and predict according to $\hat{y}_1 := \mathbb{E}_{\vec{\mathbf{c}} \sim \hat{p}_1} g_{\vec{\mathbf{c}}}(x_1)$, where $g_{\vec{\mathbf{c}}}(x_1)$ is defined above. Then, learning proceeds as the following sequence of trials $t = 1, \ldots, T-1$:

1. Get y_t and compute :

$$\hat{p}_{t+1}(d\vec{\mathbf{c}}) = \frac{e^{-\lambda \sum_{u=1}^t (y_u - g_{\vec{\mathbf{c}}}(x_u))^2}}{W_t} d\pi(\vec{\mathbf{c}}), \tag{4.23}$$

   where $W_t := \mathbb{E}_\pi e^{-\lambda \sum_{u=1}^t (y_u - g_{\vec{\mathbf{c}}}(x_u))^2}$ is the normalizing constant.

2. Get x_{t+1} and predict $\hat{y}_{t+1} := \mathbb{E}_{\vec{\mathbf{c}} \sim \hat{p}_{t+1}} g_{\vec{\mathbf{c}}}(x_{t+1})$.

Then, we have constructed a sequence of prediction $(\hat{y}_t)_{t=1,\ldots,T}$ which satisfies the following PAC-Bayesian bound.

Proposition 2. *For any deterministic sequence $(x_t, y_t)_{t=1}^T$, for any $p \in \mathbb{N}^d$, any $\lambda \leq 1/2M^2$ and any prior $\pi \in \Delta(\mathcal{C})$, one has :*

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\rho \in \Delta(\mathcal{C})} \left\{ \mathbb{E}_{\vec{\mathbf{c}} \sim \rho} \sum_{t=1}^T (y_t - g_{\vec{\mathbf{c}}}(x_t))^2 + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\}, \tag{4.24}$$

where $g_{\vec{\mathbf{c}}}(\cdot)$ satisfies (4.19) or (4.20) whereas $\mathcal{K}(\rho, \pi)$ denotes the Kullback-Leibler divergence between ρ and the prior π.

The bound of Proposition 2 gives a control of the cumulative loss of the sequential procedure described above. By using the square loss in the algorithm, this bound is more efficient than Proposition 1 where a variance term appears in the RHS. In the framework of this section, the loss function is λ-exp-concave for values of λ strictly less than $1/2M^2$. Moreover, Proposition 2 holds for any choice of prior π. It allows us in the sequel to choose a particular sparsity prior in order to state a sparsity regret bound of the form (4.22).

4.2.3 Sparsity regret bounds

The main motivation to introduce our prior is to promote sparsity in the following sense. In $g_{\vec{\mathbf{c}}}(\cdot)$, we want a codebook $\vec{\mathbf{c}} \in \mathbb{R}^{\sum_{j=1}^d m_j p_j}$ where $\vec{\mathbf{c}} = (\mathbf{c}_1, \ldots, \mathbf{c}_d)$ is such that $|\mathbf{c}_j|_0$ is small for any $j = 1, \ldots, d$, where :

$$|\mathbf{c}_j|_0 = \text{card}\{i_j = 1, \ldots, p_j : c_{j,i_j} = (0, \ldots, 0)^\top \in \mathbb{R}^{m_j}\}.$$

To deal with this issue, we propose a product of d group-sparsity priors according to :

$$d\pi_{S,d}(\vec{\mathbf{c}}) := \prod_{j=1}^d \prod_{i_j=1}^{p_j} \left\{ a_\tau \left(1 + \frac{|c_{j,i_j}|_2^2}{6\tau^2} \right)^{-\frac{3+m_j}{2}} \right\} d\vec{\mathbf{c}}, \tag{4.25}$$

for some constant $a_\tau > 0$. This prior consists of a product of d products of p_j multivariate Student's distribution $\sqrt{2}\tau\mathcal{T}_{m_j}(3)$, where $\tau > 0$ is a scaling parameter and $\mathcal{T}_{m_j}(3)$ is the m_j-multivariate Student with three degrees of freedom. It can be viewed as a generalization of the group-sparsity prior defined in Section 4.1 where $d = 1$.

It is important to stress that in (4.25), we don't need to threshold the prior at a given radius $R > 0$ such as in Section 4.1 (see the definition of the prior in (4.12)). This is due to the presence of the square loss with bounded outputs $y \in \mathcal{Y} \subseteq [0, M]$. A straightforward application of Lemma 10 in the bi-clustering framework gives the following lemma :

Lemma 11. *Let $p \in \mathbb{N}^d$, $\tau > 0$. Consider the prior $\pi_{S,d}$ defined in (4.25). Let $\vec{\mathbf{c}} = (\mathbf{c}_1, \ldots, \mathbf{c}_d)) \in \mathbb{R}^{\sum_{j=1}^d m_j p_j}$. Then, if we denote by $\mathrm{p}_{0,d}$ the translated version of $\pi_{S,d}$ with mean $\vec{\mathbf{c}}$, we have :*

$$\mathcal{K}(\mathrm{p}_{0,d}, \pi_{S,d}) \leq \sum_{j=1}^d \left\{ (3 + m_j)|\mathbf{c}_j|_0 \log \left(1 + \frac{\sum_{i_j=1}^{p_j} |c_{j,i_j}|_2}{\sqrt{6}\tau |\mathbf{c}_j|_0} \right) \right\}.$$

In this paragraph, we state the main results of this section, i.e. sparsity regret bounds of the form (4.22) for the algorithm described in Section 4.2.2. The first result is a direct consequence of Proposition 2 and the introduction of the sparsity prior (4.25).

Theorem 20. *For any deterministic sequence $(x_t, y_t)_{t=1}^T$, let us consider the algorithm of Section 4.2.2 using prior $\pi_{S,d}$ defined in (4.25) with $\tau = \delta\{\sqrt{24Mp_+T}\}^{-1}$ for some $\delta > 0$, $\lambda = 1/2M^2$ and functions $g_{\vec{\mathbf{c}}}(\cdot)$ satisfy (4.19) or (4.20). Then if T is great enough, we have :*

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\vec{\mathbf{c}} \in \mathcal{C}} \left\{ \sum_{t=1}^T (y_t - g_{\vec{\mathbf{c}}}(x_t))^2 + C \sum_{j=1}^d (3 + m_j)|\mathbf{c}_j|_0 \log T \right\},$$

where $C > 0$ is a constant independent of T.

This result gives a sparsity regret bound where the penalty term is proportional to the sum of the ℓ_0-norm of the set of codebooks $\vec{\mathbf{c}} = (\mathbf{c}_1, \ldots, \mathbf{c}_d)$. The algorithm performs as well as the best compromise between fitting the data and complexity. It is important to highlight that the RHS does not depend on the sequence $(p_1, \ldots, p_d)$. Then, as in Section 4.1, we can consider large values of p_j to learn the number of clusters.

Unfortunately, this result is essentially asymptotic since it holds for large values of T. This is due to the control of the deviation of the random variable $g_{\vec{\mathbf{c}}'}(x_t)$ to $g_{\vec{\mathbf{c}}}(x_t)$ for any $t = 1, \ldots, T$ where $\mathbf{c}' \sim p_{0,d}$ with $p_{0,d}$ defined in Lemma 11. This problem is specific to the context of bi-clustering where the map $\vec{\mathbf{c}} \mapsto g_{\vec{\mathbf{c}}}(x)$ defined in (4.19) or (4.20) is not continuous.

As in Section 4.1, this algorithm in not adaptive since it depends on unknown quantities such as the constant $\delta > 0$ (see the proof in [L11] for a precise definition) and the horizon T. Adaptive algorithms could be arranged as in Section 4.1. We can also stress that as in Gerchinovitz [2013], we can avoid the boundedness assumption $\mathcal{Y} \subseteq [0, M]$. In this case, the choice of $\lambda > 0$ in the algorithm will depend on the sequence and an adaptive choice could be investigated following Gerchinovitz [2013]. We omit these considerations for concision but could be the core of a more advanced contribution.

Corollary 20 holds for a family $\{g_{\vec{\mathbf{c}}},\ \vec{\mathbf{c}} \in \mathcal{C}\}$ satisfying (4.19) or (4.20). An inspection of the proof shows that a sufficient condition for the family $\{g_{\vec{\mathbf{c}}},\ \vec{\mathbf{c}} \in \mathcal{C}\}$ is :

$$|g_{\vec{\mathbf{c}}}(x_t) - g_{\vec{\mathbf{c}}'}(x_t)| \leq M\mathbf{1}_{\exists x_u \in \{x_1, \ldots, x_t\} : f_{\vec{\mathbf{c}}}(x_u) \neq f_{\vec{\mathbf{c}}'}(x_u)} \text{ for any } \vec{\mathbf{c}}, \vec{\mathbf{c}}', \tag{4.26}$$

where $f_{\vec{\mathbf{c}}} : \prod \mathbb{R}^{m_j} \mapsto \prod \{1, \ldots, p_j\}$ is the nearest neighbor quantizer associated with the d-tensor codebook $\vec{\mathbf{c}}$. This inequality holds in particular for families (4.19) or (4.20). Corollary 7 proposes to generalize the previous regret bound to a richer class of base functions :

$$\{g_{\vec{\mathbf{c}}}^{\mathrm{k}},\ \vec{\mathbf{c}} \in \mathcal{C},\ \mathrm{k} \in \{1, \ldots, N\}\},$$

where for any value of $\mathrm{k} = 1, \ldots, N$, (4.26) holds for $g_{\vec{\mathbf{c}}}^{\mathrm{k}}$. Functions $g_{\vec{\mathbf{c}}}^{\mathrm{k}}$ includes the previous cases (4.19) and (4.20) but any other labelizer $g_{\vec{\mathbf{c}}}$ constructed thanks to the set of past observations in the cell associated with $\vec{\mathbf{c}}$ could be considered. Given this family of N labelizers, we can advance the following prior in the algorithm described above :

$$\pi_{S,d,N} d(\vec{\mathbf{c}}, \mathrm{k}) = \prod_{j=1}^{d} \prod_{i_j=1}^{p_j} \left\{ a_\tau \left(1 + \frac{|c_{j,i_j}|_2^2}{6\tau^2} \right)^{-\frac{3+m_j}{2}} \right\} d\vec{\mathbf{c}} \times \frac{1}{N} \sum_{\mathrm{k}=1}^{N} \delta_{\mathrm{k}} d\mathrm{k}. \tag{4.27}$$

The introduction of (4.27) allows to enlarge the family of decision functions. It leads to a better sparsity regret bound with an extra $\log N$ term due to the number of base labelizers :

Corollary 7. *For any deterministic sequence $(x_t, y_t)_{t=1}^T$, consider algorithm of Section 4.2.2 using prior $\pi_{S,d,N}$ defined in (4.25) with $\tau = \delta\{\sqrt{24Mp_+}T\}^{-1}$ for some $\delta > 0$, $\lambda = 1/2M^2$. Then :*

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \leq \inf_{(\vec{\mathbf{c}}, \mathrm{k}) \in \mathcal{C} \times \{1, \ldots, N\}} \left\{ \sum_{t=1}^{T} (y_t - g_{\vec{\mathbf{c}}}^{\mathrm{k}}(x_t))^2 + C \sum_{j=1}^{d} (3 + m_j)|\boldsymbol{c}_j|_0 \log T \right\} + 2M^2 \log N.$$

This result improves Corollary 20 since the infimum is the RHS could involves different indexes k. The prize to pay is an extra $M^2 \log N$ term due to the introduction of the parameter k in the algorithm. For instance, consider the case $\mathcal{Y} = [0, M]$. If $N = 2$ in Corollary 7 and the family $\{g_{\vec{\mathbf{c}}}^{\mathrm{k}},\ \vec{\mathbf{c}} \in \mathcal{C},\ \mathrm{k} \in \{1, 2\}\}$, is made of forecasters (4.19) and a median estimator, the algorithm performs as well as the best strategy between the mean and the median.

4.3 Minimax regret [L9],[L11]

4.3.1 Introduction

In Section 4.1 and Section 4.2, we have proved several sparsity regret bounds for different sequential algorithms. These bounds are stated in the worst case scenario and have shown different behaviour with respect to time horizon T. In online clustering, the sparsity regret bounds have a residual term of order $\sqrt{T}\log T$ whereas in bi-clustering, we found better rates in $\log T$. These results are not surprising since many online learning problems give rise to similar bounds, depending on the properties of the loss functions. However, in the setting of online clustering, it is natural to ask if better algorithms exist, i.e. if lower regret could be proved for these problems.

In the context of prediction with expert advices, many authors have investigated the minimax value of the game. Given a sequence $(y_t)_{t=1}^T$, and associated experts advices $\mathbf{p}_t := (p_{t,1},\dots,p_{t,N})$, Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth [1997] have focused on the absolute loss and proved a minimax value of order $O(\sqrt{T\log N})$. In Haussler, Kivinen, and Warmuth [1998], a unified treatment of the problem is suggested with a general class of loss functions. In this context of prediction with a finite - and static - set of experts, the minimax regret is given by :

$$\mathcal{V}_T(N) := \inf_{(\hat{y}_t)} \sup_{(\mathbf{p}_1,\dots,\mathbf{p}_T)} \sup_{(y_t)} \left\{ \sum_{t=1}^T \ell(\hat{y}_t, y_t) - \min_{k=1,\dots,N} \sum_{t=1}^T \ell(p_{t,k}, y_t) \right\},$$

where ℓ is a loss function. Asymptotic behaviours for $\mathcal{V}_T(N)$ when $T\to\infty$ have been stated from $\log N$ to $\sqrt{T\log N}$ depending on particular assumptions over the loss function, such as differentiability. Many examples are provided in Haussler, Kivinen, and Warmuth [1998], including the square loss, the logarithmic loss or the absolute loss.

Very often, the proofs of the lower bounds in the deterministic setting use probabilistic arguments. Surprisingly, by considering stochastic i.i.d. generating processes for the sequence of outcomes, we can achieve tight bounds that match - at least asymptotically [3] - to the upper bounds. The starting point is the following inequality :

$$\mathcal{V}_T(N) \geq \inf_{(\hat{y}_t)} \mathbb{E}_{P^{N\times T}} \mathbb{E}_{Q^T} \left\{ \sum_{t=1}^T \ell(\hat{y}_t, Y_t) - \min_{k=1,\dots,N} \sum_{t=1}^T \ell(p_{t,k}, Y_t) \right\},$$

where $p_{t,k}$ are i.i.d. from P and $Y_1,\dots,Y_T$ are i.i.d. from Q. The rest of the proof consists in finding particular measures P and Q in order to maximize the lower bound. In this section, we want to state the same kind of result in the context of online clustering. Using simple probabilistic tools, we prove minimax results in the context of online clustering and bi-clustering.

4.3.2 Minimax regret in online clustering

According to Section 4.1, we want to investigate the optimality of Theorem 15. For this purpose, we introduce in the sequel the following assumption :

Sparsity assumption $\mathcal{H}(s)$: *Let $R>0$ and $T\in\mathbb{N}^*$. Then, there exists a sparsity index $s\in\mathbb{N}^*$ such that $|\boldsymbol{c}^\star_{T,R}|_0 = s$, where :*

$$\boldsymbol{c}^\star_{T,R} := \arg\min_{\boldsymbol{c}\in\mathcal{B}^T(R)} \left\{ \textstyle\sum_{t=1}^T \ell(\boldsymbol{c}, x_t) + |\boldsymbol{c}|_0 \sqrt{T}\log T \right\},$$

where $\mathcal{B}^T(R) = \{\boldsymbol{c} = (c_1,\dots,c_T) : |c_j|_2 \leq R, \forall j\}$.

This sparsity assumption is related with the structure of the individual sequence $x_t, t=1,\dots,T$. It means that the sequence could be well-approximated by s codepoints since the infimum is reached for a sparse codebook $\mathbf{c}^\star_{T,R}$. In what follows, we also introduce the set :

$$\omega_{s,R} := \left\{ (x_t)_{t=1}^T \text{ such that } \mathcal{H}(s) \text{ holds} \right\} \subseteq \mathbb{R}^{dT}.$$

With this notation, we have shown essentially that for any $s\in\mathbb{N}^*$, any $R>0$, the online algorithm presented in Section 4.1 satisfies :

$$\sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^T \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathcal{B}^T(R):|\mathbf{c}|_0=s} \sum_{t=1}^T \ell(\mathbf{c}, x_t) \right\} \leq \text{const.} \times s\sqrt{T}\log T.$$

[3] More recently, Audibert [2009] has given non-asymptotic lower bounds in both statistical and online learning by using the same probabilistic reduction scheme.

Then, for any $s \in \mathbb{N}^*$, $R > 0$ we could investigate a lower bound according to :

$$\inf_{(\hat{\mathbf{c}}_t)} \sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^T \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathcal{B}^T(R):|\mathbf{c}|_0=s} \sum_{t=1}^T \ell(\mathbf{c}, x_t) \right\} \geq \text{const.}' \times s\sqrt{T}\log T.$$

Unfortunately, in the inequality above, the infimum in taken over any $(\hat{\mathbf{c}}_t)_{t=1}^T$, that is with no restriction with respect to the ℓ_0-norm. Then, the LHS could be arbitrarely small and the lower bound does not match with the upper bound of Section 4.1. To impose a sparsity assumption for $(\hat{\mathbf{c}}_t)$, we need to introduced a penalized loss. Next theorem provides a lower bound for an augmented value $\mathcal{V}_T(s)$ defined as :

$$\mathcal{V}_T(s) := \inf_{(\hat{\mathbf{c}}_t)} \sup_{(x_t)\in\omega_{s,R}} \left\{ \sum_{t=1}^T \left(\ell(\hat{\mathbf{c}}_t, x_t) + \frac{\log T}{\sqrt{T}} |\hat{\mathbf{c}}_t|_0 \right) - \inf_{\mathbf{c}\in\mathcal{B}^T(R):|\mathbf{c}|_0=s} \sum_{t=1}^T \ell(\mathbf{c}, x_t) \right\}, \tag{4.28}$$

In (4.28), we add a penalization term for each $\hat{\mathbf{c}}_t$, in terms of ℓ_0-norm. As a result, to capture the asymptotic behaviour of $\mathcal{V}_T(s)$, we also need to state an upper bound with a penalized loss as in (4.28). This is done in the following theorem that combines an upper and lower bound for the minimax regret.

Theorem 21. *Let $s \in \mathbb{N}^*$, $R > 0$ such that :*

$$s \leq \left\lfloor \frac{3}{2} \left(\frac{R\sqrt{T}}{14\log T} \right)^d \right\rfloor. \tag{4.29}$$

Then :

$$s\sqrt{T}\log T(1 + o_T(1)) \leq \mathcal{V}_T(s) \leq s\sqrt{T}(\log T)^2. \tag{4.30}$$

The result of Theorem 21 gives the order of (4.28), up to a $\log T$ term. Assumption (4.29) over s is necessary for the statement of the lower bound. A similar hypothesis is used in Bartlett, Linder, and Lugosi [1998]. It is necessary here to construct the family of measures in the probabilistic reduction scheme described above.

Proof: The proof of the first inequality is based on the probabilistic method described above, where we replace the supremum over the individual sequence in $\mathcal{V}_T(s)$ by an expectation. Gathering with a suitable choice of the probability distribution inspired from Bartlett, Linder, and Lugosi [1998], we conclude the proof.

To prove the second inequality, we can use Proposition 1 to the penalized loss function $\ell_\alpha(\mathbf{c}, x) = \ell(\mathbf{c}, x) + \alpha|\mathbf{c}|_0$ with $\alpha = \log T/\sqrt{T}$ to get :

$$\sum_{t=1}^T \ell_\alpha(\hat{\mathbf{c}}_t, x_t) \leq \inf_{\rho\in\Delta(\mathbb{R}^{dp})} \left\{ \mathbb{E}_{\vec{\mathbf{c}}\sim\rho} \sum_{t=1}^T \ell_\alpha(\mathbf{c}, x_t) + \frac{\mathcal{K}(\rho,\pi)}{\lambda} + \frac{\lambda}{2} \mathbb{E}_{(\hat{p}_1,\dots,\hat{p}_T)} \mathbb{E}_{\vec{\mathbf{c}}\sim\rho} \sum_{t=1}^T [\ell_\alpha(\mathbf{c}, x_t) - \ell_\alpha(\hat{\mathbf{c}}_t, x_t)]^2 \right\}.$$

Applying the same paths as in the proof of Theorem 15, a choice of $\alpha = \log T/\sqrt{T}$, $\lambda = 1/\sqrt{T}$, a prior π_S with scale parameter $\tau = 1/\sqrt{T}$ and $p = \sqrt{T}$ allows to get the desired upper bound. ∎

4.3.3 Minimax regret in online bi-clustering

In the context of Section 4.2, we want to prove the minimax optimality of Theorem 20. For this purpose, we introduce a modified sparsity assumption related to the bi-clustering problem :

Sparsity assumption $\mathcal{H}'(s)$: *There exists a sparsity index $s \in \mathbb{N}^*$ such that $\sum_{j=1}^d |\mathbf{c}^\star_{T,j}|_0 = s$ where :*

$$\vec{\mathbf{c}}^\star_T := \arg\min_{\vec{\mathbf{c}}} \left\{ \sum_{t=1}^T (y_t - g_{\vec{\mathbf{c}}}(x_t))^2 + \sum_{j=1}^d (3 + m_j)|\mathbf{c}_j|_0 \log T \right\}.$$

This sparsity assumption is related with the individual sequence $(x_t, y_t), t = 1, \dots, T$. Loosely speaking, under $\mathcal{H}'(s)$, the individual sequence is made of a small number of clusters of inputs with same labels. In what follows, we also introduce :

$$\omega'_s := \left\{(x_t, y_t)_{t=1}^T \text{ such that } \mathcal{H}'(s) \text{ holds}\right\}.$$

With this notation, we have shown in Theorem 20 the existence of a sequential algorithm $(\hat{y}_t)_{t=1}^T$ such that for any $s \in \mathbb{N}^*$, for T great enough :

$$\sup_{(x_t,y_t)\in\omega'_s} \left\{ \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \inf_{\vec{\mathbf{c}}\in\mathcal{B}_1(s)} \sum_{t=1}^T (y_t - g_{\vec{\mathbf{c}}}(x_t))^2 \right\} \leq \text{const.} \times s \log T,$$

where $\mathcal{B}_1(s) := \{\vec{\mathbf{c}} : \sum_{j=1}^d |\mathbf{c}_j|_0 = s\}$. Then, for any $s \in \mathbb{N}^*$, we investigate the order of the minimax value :

$$\mathcal{V}'_T(s) = \inf_{(\hat{y}_t)} \sup_{(x_t,y_t)\in\omega'_s} \left\{ \sum_{t=1}^T (\hat{y}_t - y_t)^2 - \inf_{\vec{\mathbf{c}}\in\mathcal{B}_1(s)} \sum_{t=1}^T (y_t - g_{\vec{\mathbf{c}}}(x_t))^2 \right\}.$$

Following the guiding thread presented above, in this case we can move to a simple probabilistic setting as follows :

$$\mathcal{V}'_T(s) \geq \inf_{(\hat{y}_t)} \mathbb{E}_{\mu^T} \left\{ \sum_{t=1}^T (\hat{y}_t - Y_t) - \mathbb{E}_{\nu^N} \min_{k=1,\dots,N} \sum_{t=1}^T (Y_t - g_{\vec{\mathbf{c}}_k}(X_t) \right\},$$

where $(X_t, Y_t)_{t=1}^T \sim \mu_T$ and $(\vec{\mathbf{c}}_k)_{k=1}^N \sim \nu^N$. By choosing μ_T and ν_N maximizing the RHS, one gets :

Theorem 22. *Suppose $\mathcal{Y} = \{0, 1\}$ for simplicity and $g_{\vec{\mathbf{c}}}(\cdot)$ satisfies (4.19). Then, there exists a constant $a > 0$ such that for T large enough :*

$$\mathcal{V}'_T(s) \geq a \log T.$$

The proof of the lower bound follows Haussler, Kivinen, and Warmuth [1998]. However, since the family $g_{\vec{\mathbf{c}}}(\cdot)$ depends on the set of observations, we use in the proof a martingale version of the central limit theorem due to Brown [1971].

Nothing is more practical...

The last two decades have witnessed an increasing interest in high dimensional statistics (Bühlmann and van de Geer [2011]). Motivated by applications, many authors have studied models where the number of parameters p is larger than the number of observations n. In such a setting, two different issues have been opposed : theoretical guarantees and computational aspects. The lasso has been introduced and extensively studied to move to computational feasible algorithms. By the way, these considerations show rather well a common interest in algorithms for both - statistical and machine learning - communities and make the frontier between learning and statistics difficult to mark out (and then Breiman's point of view in Breiman [2001] disputable). Mathematical statistics and learning theory have a common motivation : improving the knowledge after observing data. This is exactly the guiding thread of the last chapter of this work, regarding concrete applications.

Each problem adressed in Chapter 5 corresponds to a collaboration with other scientists in different fields, such as biology, medicine or industry. We present three different applications that deal with plant architecture, medical diagnostic and sports analytics. The only common feature is the introduction of well-known machine learning tools, such as kernel principal component analysis, aggregation or Support Vector Machines. It shows one more time that learning theory is an awesome source of potential applications.

... than a good theory !

Chapitre 5

Machine learning for real-world problems

In this chapter, we outline some collaborations in applied statistics since my nomination in Angers. Applied statistics means that we are facing real-world problems. Each section corresponds to a collaboration with scientists from biology, medicine or industry and do not fit exactly the setting of Chapter 2-3-4. Consequently, this chapter is not directly bind with these theoretical results. Nevertheless, we use tools at the heart of learning theory such as kernel principal components analysis, aggregation with exponential weights or Support Vector Machines.

5.1 QTL mapping with kernel methods [L5],[L17]

5.1.1 Introduction

'Plant architecture' refers to spatial and topological structure of plants (Barthélémy and Caraglio [2007]) and determines important aspects of plant function, including productivity (Sakamoto and Matsuoka [2004]), mechanical stability (Niklas [1994]), leaf-display efficiency (Pearcy, Muraoka, and Valladares [2005]), and disease resistance (A., Milbourne, Ramsay, Meyer, Chatot-Balandras, Oberhagemann, Jong, Gebhardt, Bonnel, and Waugh [1999]). Therefore, phenotyping method of plant architecture is necessary for (i) understanding the relationship between plant form and function, (ii) the genetic improvement of crop plants, as well as (iii) the development of simulation models of plant growth. Previous studies have thus developed various methodologies for phenotyping plant architectures, such as topological (Godin and Caraglio [1998], Ferraro and Godin [2000], Segura, Ouangraoua, Ferraro, and Costes [2008]), three-dimensional (Pearcy, Muraoka, and Valladares [2005], Godin, Costes, and Sinoquet [1999]), allometric (Niklas [1994]), fractal (Ferraro, Godin, and Prusinkiewicz [2005]), and stochastic approach (Guédon, Barthélémy, Caraglio, and Costes [2001], Costes and Guédon [2002], Renton, Guédon, Godin, and Costes [2006]). These approaches have successfully analyzed and modeled precise plant architecture and its development. However, few studies apply them for phenotyping a large number of plants, which is required in the studies on genetic mapping of Quantitative Trait Loci (QTL) controlling plant architecture.

The QTL mapping of plant architecture is a critical step for understanding the genetic determinism of plant architecture and its genetic improvement by molecular breeding (Sakamoto and Matsuoka [2004]), but it requires phenotyping a large number of plants ($n > 100$). The elaborated methodologies of phenotyping plant architecture are quite labor-intensive and are not applicable to a large number of plants. As plant architectural traits are continuous and are likely to change with changes in environmental conditions (i.e., low heritability) (Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011]), replicated phenotypic measurements are necessary to evaluate their genetic variances. Furthermore, QTL analyses of phenotypic data measured at a single time point are too simple to reveal the genetic control of developmental processes of plant architecture. The functional mapping approach that fits mathematical models on growth trajectories and analyses genetic determinants of the model parameters is necessary to elucidate the genetic and developmental basis of plant growth and structure (Ma, Casella, and Wu

[2002], Wu, Cao, Huang, Wang, Gai, and Vallejos [2011]). This approach requires repeated phenotypic measurements during the development as well as appropriate mathematical models. These difficulties of phenotyping plant growth and architecture were referred to as 'phenotyping bottleneck' in plants (Furbank and Tester [2011]). This phenotyping bottleneck can now be addressed by combining novel technologies such as digital imaging, spectroscopy, robotics, and high-performance computing (Furbank and Tester [2011]), while most previous studies on QTL mapping of plant growth and architecture have performed manual phenotyping using simple geometric and topological measurements, such as plant biomass, height, shoot length, diameter, branching intensity, leaf length, and width (Wu [1998], Segura, Cilas, Laurens, and Costes [2006], Upadyayula, Wassom, Bohn, and Rocheford [2006], Onishi, Horiuchi, Ishigoh-Oka, Takagi, Ichikawa, Maruoka, and Sano [2007], Segura, Ouangraoua, Ferraro, and Costes [2008], Song and Zhang [2009], Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011], Tian, Bradbury, Brown, Hung, Sun, Flint-Garcia, Rocheford, McMullen, Holland, and Buckler [2011], Zhang, Jiang, Chen, Chen, and Fang [2012]), and have analyzed them one by one. However, many of these quantitative traits were generally correlated to each other, which give rise to statistical problem in the detection of QTL.

One-by-one QTL analysis and multiple QTL Mapping

Statistical methods for detecting QTL were originally designed for a trait-by-trait study, mostly using maximum likelihood (see Lander and Botstein [1989]) or linearised approximation (see Haley and Knott [1992]). In a first study with I.N.R.A., we analyse the phenotypic data of the date of flowering over 8 years and the number of petals over two years in two populations which have the same male parent. We identify QTLs controlling these characters in [L16]. Several authors have tried to analyze complex phenotype expression such as architecture of inflorescence. For instance Upadyayula, Wassom, Bohn, and Rocheford [2006] proposes to study maize tassel inflorescence considering 13 correlated inflorescence traits whereas Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011] studies a F1 diploid garden rose population with 10 traits associated with the developmental timing and architecture of the inflorescence and with flower production. Then, a one-by-one QTL analysis is proposed to explain the continuous variation of each trait separately. The results of such an approach often suggest that several traits are influenced by the same or linked loci. From the biological viewpoint, many questions involve the interaction between multiple traits and as a result a separate one-by-one analysis is not the most efficient. Moreover, from statistical viewpoint, the power of hypothesis tests (such as QTL mapping) tends to be lower for separate analyses.

To answer to this issue, several multiple QTL mapping have been proposed in the last decade, sometimes derived from single trait methods. Jiang and Zeng [1995] suggests multiple QTL mapping to combine several traits in an unified analysis. It has the advantage to test a number of biological hypotheses concerning the nature of genetic correlation (pleiotropy, QTL× environment interaction). This method is shown to be more efficient compared with single trait analysis (see Hackett, Meyer, and Thomas [2001]) but suffers from the curse of dimensionality. The number of parameters to estimate is higher and limits statistical power and computing time. Another approach to deal with multiple traits is based on standard multivariate analysis such as Principal Component Analysis (PCA for short in the sequel). Originally used in Weller, Wiggans, Vanraden, and Ron [1996] for dairy cow data, there are based on a linear combination of the traits in which most of the information is summarised (called the Principal Component). Then, a single trait analysis can be performed on this first PC. The PCA was applied to the QTL mapping of leaf morphology (Langlade, Feng, Dransfield, Copsey, Hanna, Thébaud, Bangham, Hudson, and Coen [2005]). The coordinates of 19 points along the leaf margin and mid-vein were obtained from leaf images as a numerical summary of the shape and size of the leaf. The coordinate values were then integrated into three orthogonal axes through PCA, and QTL analysis was performed to the PC values. Another attempt is presented in the analysis of maize inflorescence architecture (Upadyayula, Wassom, Bohn, and Rocheford [2006]), giving interesting and promising results. The existence of "PC exclusive QTL" illustrates quiet well the necessity to deal with such a multivariate analysis. Gibert and LeRoy [2003] proposes extensive simulations to compare such multitraits methods, including another multivariate analysis called discriminant analysis (see Mardia, Kent, and Bibby [1979]).

Kernel methods

Biology is facing many machine learning challenges. Massive amounts of data are generated, characterized by structured and heteregeneous data (sequences, 3D structures, graphs, networks, SNP) in large quantities and high dimension. At the core of the machine learning methodology, kernel methods have been extensively used to solve many biological problem in the last two decades. We can mention for instance predictive methods for protein function annotation (Zhou, Chen, Li, and Zhou [2007]), gene expression analysis or gene selection for Microarray Data (Scholkopf, Tsuda, and Vert [2007] which surveys the topic of using kernel methods to study biological data). A striking example of a kernel method is the Support Vector Machines (SVM) algorithm due to the pioneering's works of Vladimir Vapnik. The idea of many kernel methods is to map the dataset into an infinite dimensional space, called *feature space*, where the analysis takes place. This mapping is performed by using a so-called kernel function, which measures the similarities between two inputs x and y with the value $k(x, y)$. The construction of various type of kernels, for various type of data, allow to treat many biological problem of pattern recognition, regression estimation or PCA. This idea was originally used for classification with Support Vector Machines (see Boser, Guyon, and Vapnik [1992]), or in principal component analysis with Kernel Principal Component Analysis (KPCA, see Schölkopf and Smola [2002]).

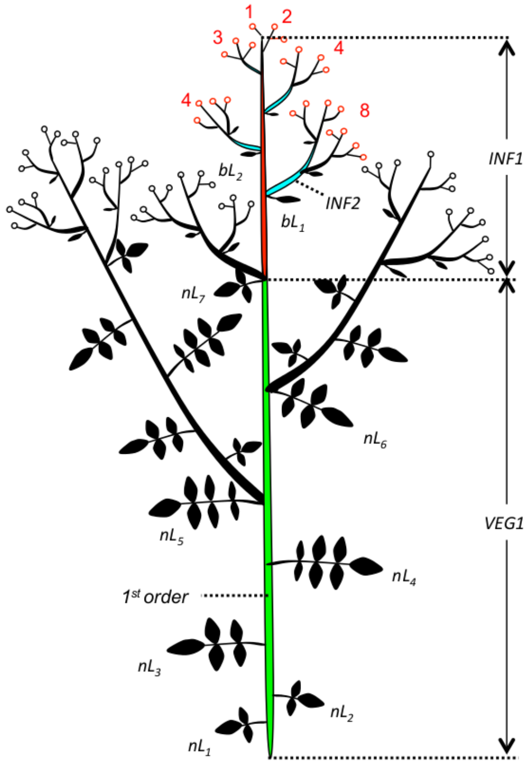
Our contribution

In this section, we aim to test the applicability of kernel PCA to QTL mapping of complex plant architectural traits. The main tools developed in this section could be use to tackle the general problem of QTL mapping of complex (sequences, 3D structure, graphs) phenotypic traits. The idea is to consider these observations directly as inputs and to work in an infinite dimensional feature space thanks to a kernel function. Kernel PCA gives a new and concise representation of the data, which is then used to perform QTL mapping without the problem of multiple, correlated data.

Specifically, we apply the method to the QTL mapping of rose inflorescence architecture. In the previous work (Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011]), QTL mapping of inflorescence architectural traits was performed in a garden rose population. In the population, roses formed a wide variety of inflorescence architecture ; a simple inflorescence formed one terminal flower and a few lateral flowers, whereas in a compound inflorescence, lateral shoots continuously branched into higher order shoots and produce numerous flowers (up to 200 flowers). We analyzed total nine traits associated with the length, node number, and branching intensity of inflorescence shoot (see inflorescence architectural traits, Figure 5.1) and found that most of these nine traits were strongly correlated to each other and they shared QTLs. We finally identified total six common QTLs (cQTLs) as genetic determinants of these nine architectural traits (see cQTL controlling the traits, Figure 5.1). In the present chapter, we use the same rose population and genetic map (Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011]) and perform QTL analysis of KPCA scores derived from a simple sequence data of flower distribution along inflorescence shoot. We hypothesize that the KPCA approach identifies a "new QTL", which was not detected by the previous work. Note that a test study using artificial data of simulated inflorescences with different types of flower distribution has been also performed in [L5] to show the ability of kernel methods to classify different inflorescence architectura. We omit these results to focus on the real dataset in this chapter dedicated to real-world applications.

Real dataset of rose population

Real dataset of inflorescence architecture was collected from the *F1* hybrid population of rose (Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011]). This population consists of a progeny of 98 diploid *F1* hybrids from a cross between diploid roses TF x RW. The female parent TF is a commercial cultivar, *The Fairy*, and the male parent RW is a hybrid of *Rosa wichurana*. Both parents develop a highly branched compound inflorescence, and their *F1* hybrids show a large genetic variation of inflorescence architecture. Three replicated clones were created for each 100 genotypes (= 98 *F1* hybrid and their parents) by vegetative propagation. A total 300 plants are cultivated in a field of INRA, Angers, France, since 2004. We collected inflorescence data from the 1st order shoot that developed in spring during two years 2008-2009. *Inflorescence* is defined as the top of the 1st order shoot that bore



Definition

nL	Normal leaf with at least 3 leaflets
bL	Bract-like leaf with 1-2 small inmature leaflets
VEG1	Vegetative part of 1st order shoot with *nLs*
INF1	Inflorescence part of 1st order shoot with *bLs*
INF2	The longest 2nd order shoot which develops from *INF1*

Inflorescence architectural traits

NV1	Number of nodes on *VEG1*
NF1	Number of nodes on *INF1*
NF2	Number of nodes on *INF2*
LV1	Average length of internodes of *VEG1* (= Length of *VEG1* / *NV1*)
LF1	Average length of internodes of *INF1* (= Length of *INF1* / *NF1*)
LF2	Average length of internodes of *INF2* (= Length of *INF2* / *NF2*)
NBF2	Number of 3rd order shoots on *INF2*
BIF2	% of lateral meristems, which develop into 3rd order shoots (= 100**NBF2* / *NF2*)
FLW	Total number of flower produced by *INF1*

Common QTL (cQTL) controlling the traits

NV1	*cQTL1*	*cQTL2*	***cQTL3***		*cQTL5*	
NF1	***cQTL1***	*cQTL2*	***cQTL3***			
NF2	***cQTL1***		***cQTL3***			
LV1	*cQTL1*			***cQTL4***	*cQTL5*	*cQTL7*
LF1			*cQTL3*	*cQTL4*	*cQTL5*	
LF2			*cQTL3*	***cQTL4***	*cQTL5*	
NBF2	*cQTL1*		***cQTL3***	*cQTL4*	*cQTL5*	
BIF2			***cQTL3***	***cQTL4***		
FLW	*cQTL1*		*cQTL3*	***cQTL4***		

Major QTLs that explained > 20% of phenotypic variance are underlined & bold.

Figure 5.1 Pictorial representation of branching structure of 1st order shoot and inflorescence garden rose. Definitions of terms are on the right. Open circle indicates a flower. The main axis corresponds to the 1st order shoot, and the lateral shoots developing from the 1st order axis are 2nd order shoots. The boundary between vegetative part (*VEG1*) and the inflorescence (*INF1*) of the 1st order shoot is defined according to the changes in leaf morphology from normal leaves (nLs) to bract-like leaves (bLs). The numbers of flower produced by 2nd order shoots are counted along *INF1* axis from the base (8, 4, 4, 3, 2, 1) and are analysed by kernel method as a vector. Other architectural trait values of the picture are as follows; *NV1* = 7, *NF1* = 6, *NF2* = 4, *NBF2* = 3, *BIF2* = 75, *FLW* = 22. Common QTL regions (cQTL) controlling these traits are also listed on the right. After modification of Figure 1 from Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011].

bract-like leaves (*INF1*, Figure 5.1). For each of the 2nd order shoots that developed from the *INF1*, we count total number of flowers. Then, we define the real dataset as a sequence of flower number per node from the base to tip of the *INF1*. The sum of them corresponds to the total number of flower per inflorescence. Measurements are made on three vigorous shoots per plant in each of the two years, and in total the data of 1460 shoots are obtained and analyzed.

5.1.2 Kernel PCA and QTL analysis

Kernel Principal Component Analysis (KPCA)

One of the most fundamental steps in data analysis and dimensionality reduction consists in approximating a given data set by a low-dimensional subspace, which is clasically achieved via Principal Component Analysis (PCA). Kernel PCA is the kernelized version of the classical PCA. Given a $(n \times p)$-matrix $X = [X_1 \dots X_p]$ of n observations $x_1, \dots, x_n \in \mathbb{R}^p$, the key step for PCA is the diagonalization of the correlation matrix, given by the inner product $\langle X_i, X_j \rangle$ between variables. Another way of expressing PCA is to consider the diagonalization of the inner product or Gram matrix XX^t, defined as :

$$(XX^t)_{ij} = \langle x_i, x_j \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the usual scalar product in the Euclidean space $\mathbb{R}^p$. In this case, principal components are calculated from the Gram matrix associated to the usual scalar product. Kernel PCA simply mimics this procedure, replacing the inner product matrix by the Gram matrix K given by :

$$K_{ij} = k(x_i, x_j),$$

where k is a *kernel function*. For each pair (x_i, x_j), the quantity $k(x_i, x_j)$ measures the similarity between x_i and x_j. From the mathematical viewpoint, k is a symmetric and positive definite function (see Cristianini and Shawe-Taylor [2000]). As a result, given a kernel function k, KPCA method provides the best linear combination of the feature variable $k(x_i, \cdot)$, where the dispersion is measured through a kernel function k. Consider a vector $x \in \mathbb{R}^p$ as a shoot, where each coordinate corresponds to the number of flowers in a node. In this framework, a kernel is a symmetric and positive definite function which associated to each pairs of shoots (x, y) a similarity measure given by $k(x, y)$. In this work, gaussian-type kernels are considered :

$$k(x, y) = \exp(-\sigma \|x - y\|^2), \tag{5.1}$$

where $\|x - y\|$ stands for a particular distance between x and y and $\sigma > 0$ is a tuning parameter. In the sequel, we called *kdist*[1] the gaussian kernel (5.1) where $\| \cdot \|$ is the standard Euclidean distance in $\mathbb{R}^p$ between shoots :

$$\text{kdist}(x, y) = \exp\left(-\sigma \sum_{i=1}^{p} (x_i - y_i)^2 \right). \tag{5.2}$$

The gaussian kernel (5.2) has a tuning parameter, namely the so-called bandwidth $\sigma > 0$. This parameter has to be chosen in a careful way. We test different values for σ from $\sigma = 0.001$ to $\sigma = 10$ according to a genetic criterion called heritability.

QTL analysis of Kernel Principal Components (KPCs)

Least square means (LS means) is computed for each KPCs for each genotype. QTL analyses are carried out using MAPQTL® 5.0 (Ooijen [2004]) on the LS means and integrated map constructed by Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011]. First, we use Kruskal-Wallis test for the rough estimation of QTL location over all KPCs derived from different kernel functions. The test ranks all genotypes according to the LS means, while it classifies them according to their marker genotype. A segregating QTL linked closely to the tested marker will result in large differences in average rank of the marker genotype classes. Based on the genetic map distribution of the significant markers, we estimate the location of underlying QTL. The linkage group with a segregating QTL must reveal a gradient in the test statistic towards the marker with the closest linkage to the QTL.

Secondly, interval mapping is performed for some KPCs in order to confirm the results of Kruskal-Wallis test and to make a more precise estimation of QTL location and effects. A LOD threshold from which a QTL is declared significant is determined according to an error rate of 0.05 over 1000 permutations of the data (Churchill and Doerge [1994]). Then, interval mapping analysis is performed with a step size of 1 cM to find regions with potential QTL effects, i.e., where the LOD score is greater than the threshold. In the region of the potential QTLs, the markers with the highest LOD values are taken as cofactors. A backward elimination procedure is used to select cofactors significantly associated with each trait at p-value< 0.02. Subsequently, multiple QTL mapping (MQM, Jansen and Stam [1994]) is performed with a step size of 1 cM. If LOD scores in the region of the potential QTLs are below the significance threshold, their cofactor loci is removed and MQM mapping is repeated. QTL positions is assigned to local LOD score maxima. Confidence intervals of the map position is indicated in centimorgans corresponding to a 1 or 2-LOD interval. The percentage of phenotypic variance explained by each QTL (r^2) is taken from the MQM mapping output. The total percentage of phenotypic variance explained by

[1] In the full version of this work, we also consider other Gaussian kernels constructed with different distances such as the kernel *kdistderiv*, associated with the discrete derivative of a shoot given by : $x = (0, 4, 2, 2, 1, 0, 0) \longrightarrow x' = (4, -2, 0, -1, -1, 0)$. We focus here on (5.2) because it gives satisfying result for the real-world dataset described previously.

all significant QTLs (R^2) is also calculated. The R^2 is then divided by h^2 (percentage scale) to estimate the proportion of genotypic variance explained by the QTL. Allelic effects is also estimated as described in Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011].

5.1.3 Main results

We assess the genetic variability of Kernel Principal Components (KPCs) derived from real datasets since QTL analysis assumes the high genetic variability of trait. Genetic variability of KPCs is evaluated by calculating its broad-sense heritability. Coarselly speaking, the total variance of KPCs is decomposed into different components. Broad sense heritability (h^2) based on genotypic mean values averaged across years is calculated as a ratio between the genetic variance and the total variance (see Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011]). The analysis is conducted using JMP software version 8.0 (SAS Institute, Inc., Cary, NC).

Generally, the first principal component Z1 has a large heritability compared to the other components, and there are high correlations between Z1 and total number of flowers (FLW) per inflorescence (Spearman's rank correlation coefficient > 0.6, p-value < 0.001). This suggests that in the studied population, a large part of genetic variation in inflorescence architecture is owing to the variation in flower number (i.e., size of inflorescence). Kruskal-Wallis test identify the markers that have significantly different Z1 scores between genotype classes. All of them are located in the genomic region of cQTL3 and/or cQTL4, where major QTLs for FLW were detected by the previous study (see Table 2, Table 1S in [L5]).

The other principal components of *kdist* functions have also substantial genetic variations ($h^2 > 0.5$). Kruskal-Wallis tests for these components show that most of their significant markers are located in the six cQTL regions (Table 2, 1S). These loci were previously detected by QTL mapping of inflorescence architectural traits, such as the length (LF), the node number (NF) and the branching intensity (BIF) of inflorescence shoots (Figure 5.1). This indicates that our kernel principal components derived from the data of a sequence of flower number along inflorescence shoot can integrate the architectural variations of inflorescence shoots.

Interestingly enough, Kruskal-Wallis test for the KPCs of *kdist* function with $\sigma = 0.025$ and 0.05 detects significant markers in linkage group 6, where the previous study have not detected any QTLs for inflorescence architecture. This indicates the discover of a new QTL. In order to confirm the result, we perform MQM mapping analysis on the KPCs derived from the function *kdist* with $\sigma = 0.025$.

MQM mappings for first five components derived from *kdist* function with $\sigma = 0.025$ identify total 11 QTLs, most of which have overlapping confidence intervals with known cQTL regions controlling architectural traits (Figure 5.2), indicating that the cQTL regions influence not only architectural traits but also the distribution pattern of flower number along inflorescence shoot. For the first KPC *Z1kdist*, there are two major QTLs, each of which explains more than 20 percent of phenotypic variance, in LG4 (*Z1kdist-1*) and in LG3 (*Z1kdist-2*). There is also a minor QTL (*Z1kdist-3*) in LG5. All these QTLs are localized in known cQTL regions controlling inflorescence architectural traits (Figure 5.2). In the cQTL4 region, Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011] identified major QTLs controlling the internode length ($LV1$, $LF2$), the branching intensity ($BIF2$), and the total number of flowers (FLW) of inflorescence shoots (Figure 5.1). The cQTL3 region also contains major QTLs controlling the number of nodes ($NF1$ and $NF2$) and the branching intensity ($BIF2$,$NBF2$) of inflorescence shoots (Figure 5.1). In support to the colocalization of QTLs, the KPC *Z1kdist* is highly correlated with inflorescence architectural traits, especially with the internode length ($LF2$), the branching intensity ($NBF2$, $BIF2$), and the total flower number (FLW) of inflorescence shoots (Spearman's rank correlation coefficient > 0.7, data not shown). The third and fourth KPCs *Z3kdist* and *Z4kdist* are also significantly correlated with the internode length ($LF1$, $LF2$), the branching intensity ($NBF2$, $BIF2$), and the total flower number (FLW) of inflorescence shoots. The QTLs for *Z3kdist* and *Z4kdist* are all colocalized with the known cQTL regions in LG3 (*Z3kdist-2*, *Z4kdist-2*), in LG4 (*Z3kdist-1*), or in LG7(*Z4kdist-1*) (Figure 5.2). In contrast, the fifth KPC *Z5kdist* is not significantly correlated either with the internode length, the branching intensity, or the total flower number of inflorescence shoots. It is significantly correlated with the number of nodes ($NV1$, $NF1$, $NF2$). In support to the result of this correlation analysis, *Z5kdist* have a major QTL (*Z5kdist-1*) in the cQTL1 region (Figure 5.2), where Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011] identified major QTLs

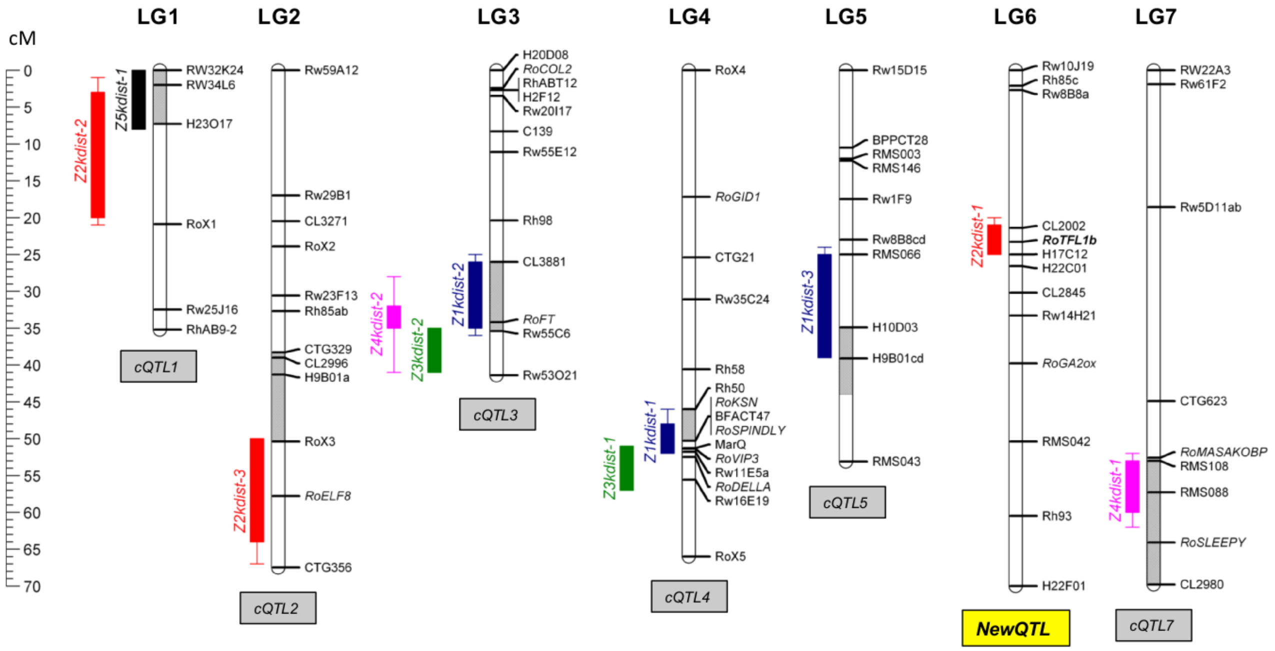


Figure 5.2 Genetic map locations of QTLs for five kernel principal components (KPCs) detected by multiple QTL mapping in 98 *F1* diploid roses derived from the cross TF x RW. The KPCs are obtained by applying kernel function *kdist* with = 0.025 to the sequence data of flower number per node along inflorescence axes. Common genomic regions of QTLs (cQTLs) for 10 inflorescence developmental traits previously identified by Kawamura, Oyant, Crespel, Thouroude, Lalanne, and Foucher [2011] are also indicated. QTLs are illustrated by boxes whose length represents the LOD-1 confidence interval. Extended lines represent the LOD-2 confidence interval. Left bar shows the map scale in cM. For QTL abbreviation, see Table 3 from [L5].

controlling the number of nodes per inflorescence shoots (Figure 5.1). Thus, the most QTLs detected for the KPCs are located in the previously identified cQTL regions controlling inflorescence architecture and are likely to be involved in the regulation of internode elongation, node production, and/or axillary branching of inflorescence shoots.

An exception is the second KPC *Z2dist*, which have a major QTL (*Z2dist-1*) in LG6 (Figure 5.2), where the previous study did not detect any QTLs. The genotypic correlation analysis shows that the *Z2kdist* is not significantly correlated with any architectural traits, such as the internode length (*LF*), the node number (*NF*) or the branching intensity (*BIF*) of inflorescence shoots. It is just weakly correlated with the total number of flower (*FLW*) per inflorescence (Spearman's rank correlation coefficient = -0.23, p-value< 0.05). Thus, the *Z2kdist* is not characterized by simple architectural traits, such as the length, number, and branching intensity of nodes. It is also not a simple measure of inflorescence size. As a result, the newly identified QTL *Z2kdist-1* might be involved with the control of flower distribution along inflorescence shoot. In the *Z2kdist-1* region, a candidate gene, *RoTFL1b*, a homologue gene of *TERMINAL FLOWER 1* of *Arabidopsis thaliana* (Iwata, Gaston, Remay, Thouroude, Jeauffre, Kawamura, Oyant, Araki, Denoyes, and Foucher [2012]), is co-localized (Figure 5.2).

5.1.4 Conclusion

This contribution demonstrates the applicability of KPCA method for QTL mapping of a complex plant architectural trait, namely the inflorescence architecture. We assess the usefulness of different kernel methods based on the calculation of heritability of KPCA components. This allows us to select the kernel function that discriminates well the genetic variance of the focused traits in the studied population. The

QTL analysis of kernel principal components identifies a new QTL, which was not detected by a trait-by-trait analysis. We have tried to characterize the function of *Z2kdist-1*. We can conjecture that the *Z2kdist* represents the distribution pattern of flower along inflorescence axis. To test the hypothesis, we can examine the correlation between the *Z2kdist* and simple indices of flower distribution along inflorescence axis. The simple indices of flower distribution are obtained by counting the number of nodes where the accumulative number of flower attains 50 percent of total number of flower. The calculations are done for each shoot both from the base and the tip of inflorescence axis (*INF1*, Fig. 5.1). The indices, obtained by counting from the base and the tip, are named as *B50* and *T50*, respectively. Either the *B50* or the *T50* are not significantly correlated with the *Z2kdist* (p-value$>$ 0.1). QTL analysis does not detect significant QTLs on LG6 either for the *B50* or for the *T50* (data not shown). Therefore, the *Z2kdist* could not be characterized by the simple indices of flower distribution tested here. A detailed pattern analysis (e.g., Guédon, Barthélémy, Caraglio, and Costes [2001]) may be necessary to interpret the *Z2kdist* and the function of *Z2kdist-1*.

The *RoTFL1b* is a candidate gene for the *Z2kdist-1*. In *Arabidopsis thaliana*, *TFL1* is expressed in shoot apical meristem to maintain meristem indeterminacy and control inflorescence architecture (Prusinkiewicz, Erasmus, Lane, Harder, and Coen [2007]). Overexpression of *TFL1* delays flower formation and forms a highly branched inflorescence, while *tfl1* mutants have a short vegetative phase and form a simple determinate inflorescence with a terminal flower (Bradley, Ratcliffe, Vincent, Carpenter, and Coen [1997]). The structure and function of *TFL1* gene is greatly conserved in plants (reviewed by McGarry and Ayre [2012]). We recently demonstrated that *RoKSN*, another *TFL1* member in rose, is expressed in shoot apical meristem and plays a role in the repression of flowering, and *ksn* mutants have a continuous flowering habit (Iwata, Gaston, Remay, Thouroude, Jeauffre, Kawamura, Oyant, Araki, Denoyes, and Foucher [2012]). Given the high degree of sequence similarity between *RoKSN* and *RoTFL1b* (Iwata, Gaston, Remay, Thouroude, Jeauffre, Kawamura, Oyant, Araki, Denoyes, and Foucher [2012]), it is likely that the *RoTFL1b* is also involved in the control of floral transition and inflorescence development in rose. Future expression analysis and physiological study will be necessary to clarify the hypothesis.

5.2 Fibrosis staging with aggregation

Fibrosis is the formation of excess fibrous connective tissue in an organ, due to a reactive process. Fibrosis can arise in many tissues within the body, such as lungs, heart, skin or intestine. In liver, cirrhosis is a result of advanced fibrosis and leads to a loss of liver function. It is most commonly caused by alcoholism, hepatitis (B and C) or other possible causes. HIFIH laboratory develops accurate and non-invasive blood-tests for identifying stage of fibrosis, for instance in non alcoholic fatty liver disease (NAFLD) (see Calès, Boursier, Chaigneau, Lainé, Sandrini, Michalak, Hubert, Dib, Oberti, Bertrais, Hunault, Cavaro-Ménard, Gallois, Deugnier, and Rousselet [2010]) or in chronic hepatitis C (Calès, Boursier, Ducancelle, Oberti, Hubert, Hunault, Lédinghen, Zarski, Salmon, and F.Lunel [2014]). In this section, we want to use aggregation methods to predict the fibrosis stage thanks to simple biomarkers in order to propose an automatic blood-test based method.

Dataset description

All of the 1012 patients included in the derivation population has a fibrosis variable F from 0 (Fibrosis absence) to 4 (cirrhosis) gathering with 6 blood-tests variables related with different quantities, levels or rates. It includes X_1 (G/l, platelets or thrombocytes), X_2 (UI/i, aspartate amino-transferase), X_3 (mmol/l, blood urea level), X_4 (%, Prothrombine blood rate), X_5 (mg/dl, Alpha2macroglobulin), X_6 (UI/l, gamma glutalyl transpeptidase quantity). Eventually, we use X_7 (Age of the patient) and X_8 (Male/Female).

Multinomial Logistic Regression (MLR)

The value of F has been measured thanks to an invasive method. In order to develop non-invasive methods based on blood-tests, we propose to use a multinomial logistic regression model from the input variables $X_1, \ldots, X_8$ described above. The logistic regression model consists in modelling the posterior probabilities $\eta_k(x) := \mathbb{P}(F = k|X = x)$, $k = 0, \ldots, K-1$ via a linear function in x. We use in the sequel the following logit transformations :

$$\forall k = 0, \ldots, K-1, \ \log\left(\frac{\eta_k(x)}{\eta_K(x)}\right) = \beta_k^\top \cdot (1, x),$$

where $(1, x) = (1, x_1, \ldots, x_8)$ for simplicity. Then, a simple calculation shows that :

$$\mathbb{P}(F = k | X = x) = \eta_k(x|\beta) = \frac{e^{\beta_k^\top \cdot (1,x)}}{1 + \sum_{j=0}^{K-1} e^{\beta_j^\top \cdot (1,x)}}.$$

The model of logistic regression is widely used in biostatistics for $K = 1$ (binary classification), where in this case there is only a single linear function. In our problem, parameters $(\beta_k)_{k=0}^{K-1}$ are usually fitted by maximum likelihood. The associated first order conditions are in matrix notation as follows :

$$X^\top (y - p) = 0,$$

where X is the data matrix with $n = 1012$ rows and $d = 8 + 1$ columns, y is the vector of fibrosis stages and p is the vector of fitted probabilities given by $(\eta_1(x_i|\beta), \ldots, \eta_k(x_i|\beta))^\top$. Then, a Newton-Raphson algorithm could be performed.

According to the health care professional, 9 logistic regressions were calculated thanks to the dataset. The first one is the multinomial logistic with $d = 9$ when we consider the entire set of feature variables. It is called $\mathrm{MLR}_{\mathrm{tot}}$. Then, we construct 8 other logistics by avoiding one variable from the dataset. It gives MLR_1, ..., MLR_8 where MLR_k is the logistic without X_k. The associated classifier are denoted as f_{tot}, and f_j, $j = 1, \ldots, 8$ and are given by the following formula :

$$f_j(x) = \arg \max_{k=0,\ldots,K-1} \eta_k(x|\hat{\beta}_j),$$

where $\hat{\beta}_j$ is the solution of the Newton-Raphson gradient descent associated with MLR_j computed with package VGAM.

Aggregation with Mirror Averaging (MA)

Aggregation methods are very popular in machine learning. The principle of the method is to construct a combination of a finite number $M \geq 1$ of base learners $\{f_1, \ldots, f_M\}$, in order to give an accurate prediction strategy. This is an alternative to the well-known empirical risk minimization principle, which selects a particular classifier in a given family. The main motivation is as follows. Very often, a particular classifier can not perform well on each occurence of a test set. Then, the use of a combination of classifiers instead of a single method can lead to better results. Most of the time, the sample is divided into two parts : the first part is used to construct a family of base learners whereas the second part is used to construct the associated weights[2].

Equipped with a family of preliminary functions, denoted as $\Phi = \{f_1, \ldots, f_M\}$, we construct our final decision sequentially. At each trial $t = 1, \ldots, n$, for $j = 1, \ldots, M$, we compute the empirical risk $r_{t,j}$ of classifier $f_j \in \Phi$ at time t and associated weights $\hat{w}_{t,j}$ as follows :

$$\hat{w}_{t,j} = \frac{e^{-\lambda r_{t,j}}}{W_t}, \text{ where } r_{t,j} = \sum_{i=1}^{t} \mathbf{1}_{Y_i \neq f_j(X_i)}, \tag{5.3}$$

whereas $W_t > 0$ is such that $\sum_{j=1}^M \hat{w}_{t,j} = 1$ and $\lambda > 0$ is a temperature parameter. Eventually, we proceed to the final step called "mirror averaging" and construct the final weights :

$$\hat{w}_j = \frac{1}{n} \sum_{t=1}^{n} \hat{w}_{t,j}. \tag{5.4}$$

We hence obtain an aggregate called mirror averaging (MA) defined as $\hat{f}_{\mathrm{MA}}(\cdot) = \sum_{j=1}^M \hat{w}_j f_j(\cdot)$.

[2] In Barron and Leung [2006], it is proved that in the context of linear regression, we can calculate the least-square projections and the associated aggregate with the same sample.

Result of the experiment

Following the aggregation scheme of Section 5.2, we divide the sample into two parts. The first part of the sample ($n_1 = 506$ patients chosen randomly) is used to construct the family of classifiers $\Phi = \{f_{\text{tot}}, f_1, \ldots, f_8\}$, where multinomial logistics are performed on this primary set of patients. Then, we use the second subsample of $n_2 = 506$ patients to construct the Mirror Averaging aggregate $\hat{f}_{\text{MA}}(\cdot)$.

The evolution of the performances of each MLR are given in Figure 5.3 below.

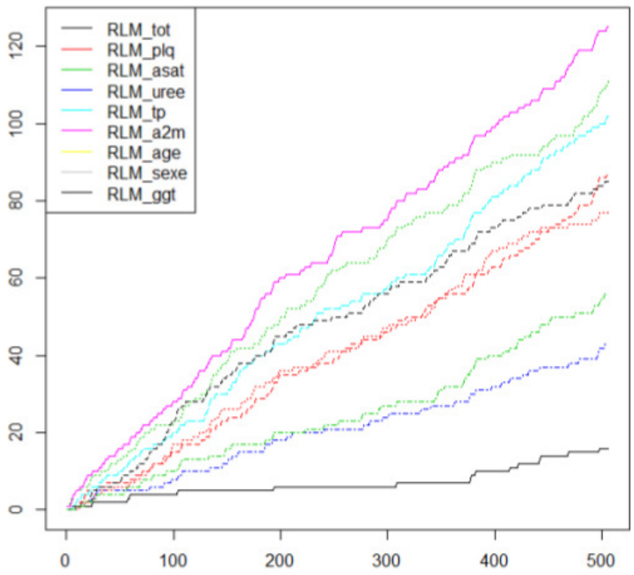
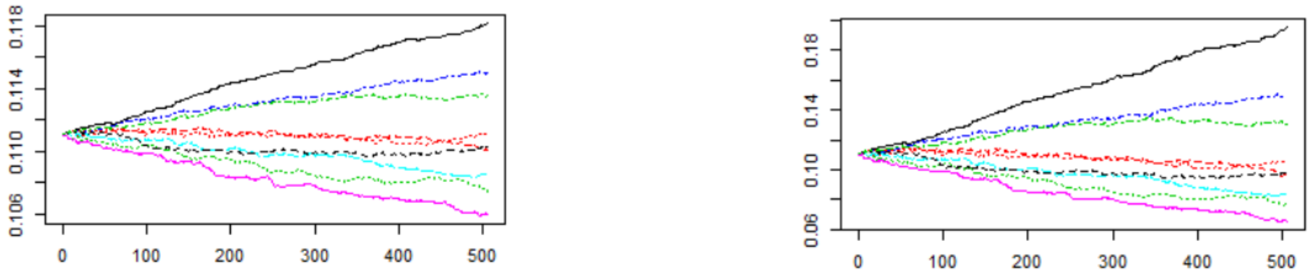


Figure 5.3 Evolution of the empirical risk $r_{t,j}$ of each $f_j \in \Phi$ over $n_2 = 506$ patients.

We can note that f_{tot}, the multinomial logistic regression based on the whole set of variables $X_1, \ldots, X_8$ has a good accuracy (3.16%) whereas other regressions give intermediate results from 8.49% for f_3 to 24.7% for f_5.

```
> erreur_rlm(reg)
   RLM_tot    RLM_plq   RLM_asat   RLM_uree     RLM_tp    RLM_a2m    RLM_age
0.03162055 0.17193676 0.21936759 0.08498024 0.20158103 0.24703557 0.16798419
  RLM_sexe    RLM_ggt
0.15217391 0.11067194
```

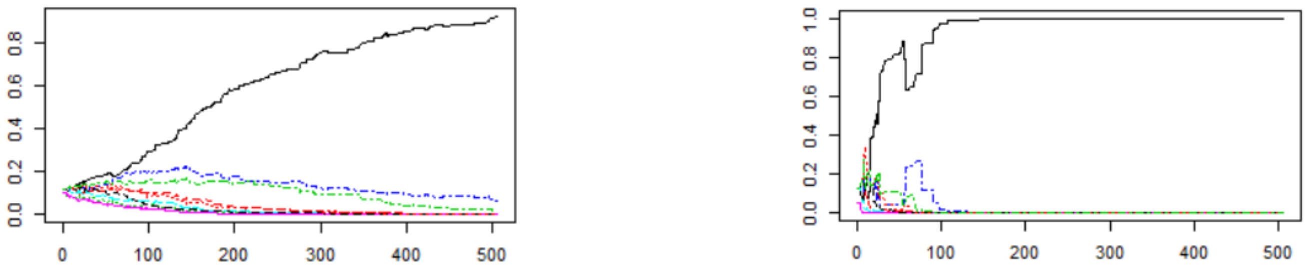
Then, we can proceed to the sequential construction of weight. Figure 5.2 below shows the evolution of the weights with small temperature parameters $\lambda = 0.001$ and $\lambda = 0.01$.



(a) $\lambda = 0.001$ (b) $\lambda = 0.01$

Figure 5.4 Evolution of weights defined in (5.3) with $n_2 = 506$ and small temperature parameters.

The influence of λ can be seen in the vertical axe. The sequence of weights can also be computed with greater temperature parameters $\lambda = 0.1$ and $\lambda = 1$.



(a) $\lambda = 0.1$ (b) $\lambda = 1$

Figure 5.5 Evolution of weights $\hat{w}_{t,j}$ defined in (5.3) with $n_2 = 506$ and large temperature parameters.

Here, the influence of f_{tot} is significantly higher. This is due to the high values of the temperature parameters. It shows rather well that by increasing the value of the temperature parameter in (5.3), we lead to an ERM strategy, where the second sample is used as a test set.

Eventually, we proceed to a leave-one-out cross validation method in order to calculate the accuracy of each aggregate $\hat{f}_{\text{MA}}$ for various temperature parameters. The result of this study shows that $\lambda = 0.38$ is the best compromise. It gives a prediction error of 2.766798% of misclassification which is detailled for each fibrosis stage in the following table :

```
F0         F1         F2         F3         F4
0.02325581 0.02040816 0.02145923 0.04000000 0.03030303
```

Conclusion

This section illustrates the power of aggregation in fibrosis staging based on blood-tests. It can be seen as a first attempt into the development of non-invasive methods with statistical learning.

5.3 Sport analytics with SVM [L13]

Sports analytics is an emerging field especially in the US (see Alamar [2013] for a survey [3]). The principle is to integrate statistics and data analysis into decision-making strategies for general managers, coaches and other professionals. It was initiated with sabermetrics (the empirical analysis of baseball) by Bill James and Nate Silver (see Silver [2007] for instance). Nowadays, analytic tools and data grow increasingly complex. One of the biggest challenges is to identify the most useful pieces of data, and then put into a player/coach decision. In this section, we propose a modest contribution to the field that deals with Basketball. We present a collaboration with ITNoveo, a small industry in Information Technology which gives rise to a smartphone application of sports analytics called Youscore.

Presentation of the smartphone application

Youscore is a mobile application designed by the general manager of ITNoveo, phd in computer science. It allows to broadcast a game play by play with a smartphone or a tablet. Watching a game, you make the score growing. Instantaneously, the scenario is available live for all the other users.

Recently, we proposed to go further. It was natural to add a statistical plug-in to use the information of the game broadcasted with Youscore. YouScorePredict (YSP) expects to predict the issue of the game. A Beta version can be downloaded on the Google Play Store here :

`https://play.google.com/store/apps/details?id=fr.youscore.android.youscorepredictplugin`

Material and methods

We have chosen in this problem a purely statistical learning point of view by collecting a database of $n = 1222$ past games of the last NBA season. The play-by-play scenario of any NBA game can be find on the ESPN website [4]. We developed a BASH script that is able to download hundreds of games based

[3] We can also refer to the blog of Nate Silver at `http://fivethirtyeight.com/`.

[4] An example can be found here : `http://scores.espn.go.com/nba/playbyplay?gameId=400489851&period=0`.

on the gameId given by ESPN. This script produced a set of 1222 HTML files, one for each game. These HTML files are nothing but structured data trees and can be represented in the memory of a computer as a DOM (Document Object Model). Exploring these raw HTML files, we can extract the evolution of the score and ignore all the other non relevant information. This has been done with DOM inspection techniques (see for instance Resig [2006]). After this process, we obtain $n = 1222$ flat plain text files, containing n lines. A given line contains 3 variables : the time of the basket, the home team score, and the visitor team score. Eventually, we come up with $\mathcal{Z}_n = \{Z_1, \dots, Z_n\}$, where each Z_i, $i = 1, \dots, n$ corresponds to the play-by-play evolution of game number i (that is the difference between the home team score and the visitor team score at each time). Then, we employ SVM to this database.

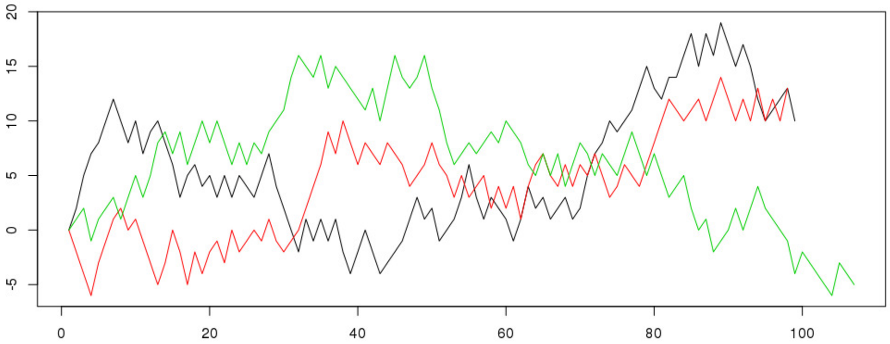


Figure 5.6 Evolution of 3 different play-by-play.

SVM is now a standard learning system based on recent advances in statistical learning theory (see Vapnik [1998]). It was originally proposed in Boser, Guyon, and Vapnik [1992] to solve the binary classification problem as follows. Consider a learning sample $\{(x_1, y_1), \dots, (x_n, y_n)\}$, where to each input $x_i \in \mathbb{R}^p$ corresponds a binary response $y_i \in \{-1, +1\}$[5]. Given an input x, it is possible to use a real-valued function $f : \mathbb{R}^p \to \mathbb{R}$ to assign the class of x : if $f(x) \geq 0$, x is supposed to be in the positive class, and otherwise to the negative class. Linear discrimination (or perceptrons) considers the case where $f(x)$ is a linear function of $x \in \mathbb{R}^p$, so it can be written as :

$$f_{w,b}(x) = \langle w, x \rangle + b, \tag{5.5}$$

where (w, b) are the parameters that control the decision rule given by $\text{sign}(f_{w,b})$. The idea of SVM is to learn from the learning sample $\{(x_1, y_1), \dots, (x_n, y_n)\}$ these parameters, by giving an hyperplane which optimally separates the two classes.

In the linear case (4), (w, b) is defined to solve the following optimization problem :

$$\left\{ \begin{array}{l} \min \|w\|^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to } y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i,\ i = 1, \dots n, \end{array} \right. \tag{5.6}$$

where $\xi_i \geq 0$ are slack variables and $C > 0$ is a regularization parameter that avoids overfitting. The unique solution of this problem gives the so-called soft margin hyperplane with geometric margin $\gamma = 1/\|w\|^2$ (the distance between the hyperplane and the nearest sample of each class).

Eventually, the kernel method of SVM is defined as a soft margin hyperplane in a high dimensional feature space, using a kernel function k as in Section 5.1. More precisely, deriving the primal Lagrangian for the optimization problem gives rise to the following objective function :

$$W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j \left(\langle x_i, x_j \rangle + \frac{1}{C} \delta_{ij} \right),$$

where δ_{ij} is the Kronecker δ defined to be 1 if $i = j$ and 0 otherwise. To move to the more general kernel version, we have to replace in $W(\alpha)$ the scalar product $\langle x_i, x_j \rangle$ by the quantity $k(x_i, x_j)$. The decision

[5] In the sequel, the output of a game x corresponds to the win ($y = 1$) or loss ($y = 0$) of the home team

rule is given by :

$$\operatorname{sign}(f(x)) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i^* k(x_i, x) + b^*\right).$$

It is equivalent to the hyperplane in the feature space implicitly defined by the kernel k. We refer the interested reader to Cristianini and Shawe-Taylor [2000] for a complete and readible introduction about SVM.

Experimental results

The problem we have at hand depends on time t when the user want to ask for a prediction. Let us fix an arbitrary time t that could be chosen by the user. Given t, we construct a training set $\mathcal{D}_n = \{(X_i, Y_i),\ i = 1, \ldots, n\}$ of $n = 1222$ games based on the previous database $\mathcal{Z}_n = \{Z_1, \ldots, Z_n\}$. Each Z_i gives the entire score evolution from the first basket to the last basket. We can then stop at time t for any given t. However, the length of this "play-by-play" evolution until time t differs since the number of baskets is not the same for two distinct games. Then, we propose to choose a bandwidth of length $h \in \mathbb{N}^*$ to look at the last h baskets. Interestingly enough, this bandwidth has to be chosen adequately. In the sequel, we use simple cross-validation methods to choose a fixed (i.e. indepent of t) value for h. A good bandwidth is around $h = 30$ baskets. It gives a vector $X_i \in \mathbb{R}^{30}$, where each dimension corresponds to a basket. Then, an associated class $Y_i \in \{-1, 1\}$ is observed for these games, where $Y_i = 0$ when the home team looses the game (and $Y_i = 1$ for a win).

We have a training set (X_i, Y_i), $i = 1, \ldots, n$ where $X_i \in \mathbb{R}^h$ and $Y_i \in \{-1, 1\}$. The SVM machinery is used with software R, using library kernlab. We use a Gaussian kernel with bandwidth selected with V-fold cross-validation, with $V = 5$. The results of the associated cross-validation errors are illustrated in Figure 5.7 below, where the algorithm gives quite good results compared with a simple - but also accurate - strategy that gives class 1 to X_i when the home team leads the score at time t, and 0 otherwise (this strategy is called "basic fan" in the sequel).

The performances of the method is illustrated for $h = 30$ and different values of t (see Figure 5.7 below). Of course, when the prediction is performed at the end of the game, the performances are very good (around 95% at origin of the horizontal axe in Figure 5.7). The performances of YSP prediction rule based on SVM on the past 30 baskets are better than the basic fan. The basic fan still predicts that the winner of the game is the winner at time t. As a result, at the end of the game, this strategy is very efficient. Nevertheless, when we go earlier in the game, our method outperforms the basic fan. It means that by looking at the 30 last baskets and performs a kernel method, we can extract more information. Of course, better results could be expected by a painstaking calibration of h. For instance, it is clear that the optimal value of h depends on time t when the prediction is performed. It could be a way of improving these results. We can also mix several strategies, such as SVM with the basic learner in order to outperform a single method.

Conclusion

As a conclusion, we have built an operational prediction rule for Basketball using androïd system. This method permits to forecast the winner of a game by loading the play-by-play score evolution live. Many advancements could be considered in the future. From the statistical viewpoint, some hyper-parameters could be calibrated in the method, such as the window h in the database. Another direction will be to advance different prediction methods such as Random Forest or possible aggregates. Another track could be to take into account several statistical informations such as the field goal percentage of several key players, or any other information (rebounds, assists, dunks and so on). Eventually, this problem could also be viewed as an online learning problem where at each time t, we want to predict the next event. In this case, we could aggregate several expert's advices (which have to be constructed) and lead to good accuracy at the end of the game. This problem is more challenging since horizon T is not very high in this case (around 100 events for a NBA basketball game).

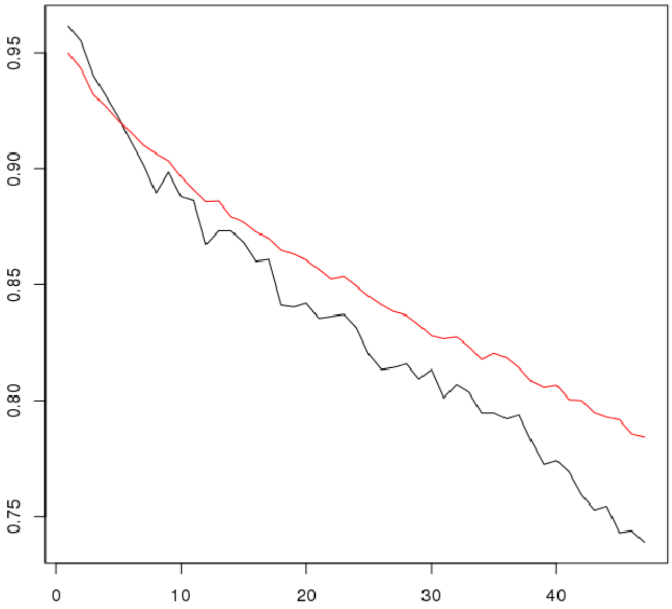


Figure 5.7 Prediction accuracy of YSP (red line) against the basic fan (black line). The horizontal axe gives time t (time left in minuts) when the prediction is performed. The basic fan still predicts that the winner of the game is the winner at time t.

Open problems

We have presented recent contributions in learning theory that I hope will shed some light into the connections between mathematical statistics and machine learning. As a conclusion, we list several open problems. Some of them are just minor extensions of the results of this habilitation whereas other ones need more investigation. By the way, the present manuscript raises many questions that could be adressed in the future.

Open problems related with ISL (Inverse Statistical Learning)

Open problem 1: At the light of Chapter 2, the study of the minimax rates of convergence in discriminant analysis is not completely satisfactory. In the plugin framework, lower bound holds for particular margin parameters $\alpha \leq 1$. The main problem with this lower bound is the opposition between the margin assumption and the noise assumption. It makes the construction of a good hypothesis family very nasty in Lemma 1. Indeed, in standard lower bounds in classification with Assouad's lemma, the marginal densities are very irregular. Here, due to the assumption on the inverse problem, we need to consider cosine type densities, as in Butucea [2007]. An open problem is to get the same result for any $\alpha \geq 0$.

Open problem 2: In the Hölder boundary case, while I'm writing this dissertation, minimax rates remain an open problem. We have spent many time with Clément Marteau to this end. From my point of view, the problem comes from the upper bound, and precisely the proposed plug-in procedure. Indeed, the regularity assumption deals with the boundary of $G_K^\star$ whereas the proposed estimators deal with the conditional densities. A possible direction for a future attempt in to try to compute the operator of inversion in our problem. In other words, what is the expression of operator A in the following equation :

$$G_\eta^\star = \{x \in \mathbb{R}^d : x_d \leq Ab^\star(x_1, \ldots, x_{d-1})\} \in \arg\min \left\{ \frac{1}{2} \left(\int_{G^C} f * \eta dQ + \int_G g * \eta dQ \right) \right\},$$

where $G^\star := \{x \in K : b^\star(x_1, \ldots, x_{d-1}) \leq x_d\}$ is the minimizer of the Bayes risk.

Open problem 3: Another point of view in inverse statistical learning would be to try to classify a new observation $Z = X + \epsilon$, in the presence of errors in variables. Is it necessary in this case to use an indirect approach with deconvolution kernel ? This question has been already investigated in test theory by Clément Marteau (see also his habilitation thesis), where direct approaches are proposed for inverse problems. In the classification setting, we need to compute the margin parameter in the presence of errors-in-variables, which is a rather difficult task, since the noise assumption is global whereas the margin is local.

Open problem 4: In this manuscript, we do not adress the problem of model selection of the hypothesis space. In Chapter 3, we spend some time to select the bandwidth in a fixed model $\mathcal{G}$, or equivalently for a fixed number of clusters k in clustering. An open problem is to offer penalization techniques. We believe that this is possible, since model selection and risk bounds use the same machinery (see for instance Massart [2007], van de Geer [2000] or Koltchinskii [2006]).

Open problem 5: In Chapter 2, we conduct the main lower and upper bounds in the context of discriminant analysis. In the direct case, Mammen and Tsybakov [1999] propose a minimax study in discriminant analysis whereas Audibert and Tsybakov [2007] study the classification context. The minimax study of this thesis could be moved to the classification setting with simple modifications.

Open problem 6: In all this study, we restrict ourselves to moderately ill-posed inverse problems, namely a polynomial decreasing of the characteristic function of the noise density (or more generally the spectrum of operator A). Extensions to exponentially decreasing cases could be done. This will deteriorate the rates exactly as in standard statistical inverse problem. More precisely, the lipschitz constant $c(\lambda)$ in Definition 2 of Chapter 2 will have an exponential behaviour. Of course, fast rates are prohibited in this case.

Open problem 7: It is quite standard in statistical inverse problem to consider a noisy operator A. In errors-in-variables model, it corresponds to an unknown noise density which has to be estimated thanks to repeated measurements. In our framework, the empirical risk will be modified by plugging an estimation of the Fourier transform of the noise. This framework could be adresses in the future from the theoretical point of view, where a more complicated empirical process theory has to be performed. This problem was considered in the simulation computation of Noisy k-means, where we add this estimation step in the algorithm thanks to an i.i.d. sample ϵ_u, $u = 1, \ldots, m$. It does not deteriorate the results, at least from a practical point of view.

Open problem 8: The construction of noisy k-means reveals an interesting phenomenon in Theorem 8. In the direct case, the k-means construction is based on :

$$\hat{\mathbf{c}}_{\ell,j} = \frac{\sum_{i=1}^n \int_{V_j} x_\ell \delta_{X_i} dx}{\sum_{i=1}^n \int_{V_j} \delta_{X_i} dx},$$

whereas the noisy k-means is defined according to :

$$\widetilde{\mathbf{c}}_{\ell,j} = \frac{\sum_{i=1}^n \int_{V_j} x_\ell \widetilde{\mathcal{K}}_h(Z_i - x) dx}{\sum_{i=1}^n \int_{V_j} \widetilde{\mathcal{K}}_h(Z_i - x) dx}.$$

To build a noisy version of the k-means, we just need to put a deconvolution kernel centered at each observations, instead of a Dirac function. A natural question is the following : can we use this simple trick to produce other noisy algorithm, such as a Noisy SVM or any other kernel method for instance ?

Open problem 9: Open problem 8 suggests to put a deconvolution kernel at each observations. In the direct case as well, we could put a standard kernel instead of a Dirac function at each observations. In this case, we smooth the minimization problem and we conjecture that the dependence on the initialization could be reduced. Unfortunately, at the same time, we add a bias in the estimation procedure. An interesting direction is to test this procedure numerically. Is there an adequate choice of the kernel (and the bandwidth) in order to trade off these two opposing phenomena (namely convexifiation and bias) ?

Open problems related with bandwidth selection

Open problem 10: The choice of the bandwidth in noisy k-means suffers from the non-convexity of the k-means loss function. Indeed, the theoretical results of Chapter 3 provides two bandwidth selection rule for noisy k-means. Unfortunately, these methods depend on the global minimizer of the deconvolution empirical risk, which is not available in practice. An open problem is to derive a data-driven selection rule which takes into account this difficulty. One could think for instance at the following procedure in the isotropic case :

$$\hat{h} = \max\left\{ h \in h_a \ : \ \widehat{R}_{h'}(\hat{\mathbf{c}}_{h,i_h^*}) - \widehat{R}_{h'}(\hat{\mathbf{c}}_{h',i_{h'}^*}) \leq 3\delta_{h'}, \ \forall h' \leq h \right\},$$

where for a given h, $\hat{\mathbf{c}}_{h,i_h^*}$ is the solution of the noisy k-means algorithm with initialization i_h^* minimizing the empirical distortion $\widehat{R}_h(\hat{\mathbf{c}}_{h,i})$, for different $i = 1, \ldots, I$ initializations.

Open problem 11: The choice of k is not adressed in this manuscript. In the direct case, there exists several methods, based on the clustering with different values of k, and a minimization (or maximization) of some criterion (see Fischer [2011] and the references therein). In the presence of noisy observations, this problem is interesting since the additional noise could hide the presence of some clusters. At the first glance, we could applied standard methods based on the observations of a distortion (see for instance the "Gap statistics" in Tibshirani, Walther, and Hastie [2001]), where we replace the direct distortion by the deconvolution distortion with a suitable deconvolution kernel.

Open problem 12: Chapter 3 proposes two different selection rules based on Lepski's contributions. In the isotropic case, we compare empirical risks instead of estimators whereas for the anisotropic case, we compare empirical gradients to estimate a bias variance decomposition. Of course, for a real hyperparameter $\lambda > 0$, we can compare empirical gradients instead of empirical risks in the standard Lepski's procedure. It might give a method easier to calibrate since it does not depend on the margin constant.

Open problem 13: A credible application of the bandwidth selection method of Chapter 3 is the problem of image denoising. It is well-known that linear estimates can degrade dramatically if the random noise obeys a non-Gaussian distribution. Then, using for instance a Huber loss, we could investigate numerically the problem of anisotropic bandwidth selection in image denoising with Local Polynomial Approximation.

Open problem 14: In the bandwidth selection method presented in Section 3.2, we restrict ourselves to a finite dimensional space $\mathbb{R}^m$, where $m \geq 1$ is the (small) dimension of the parameter space. Many statistical problems could be treated. However, a natural extension to the high dimensional setting is an appealing open problem.

Open problems related with online learning

Open problem 15: In online learning, the use of exponential weighted averages, as well as Gibbs measure, is motivated by standard PAC-Bayesian bounds such as in Mac Allester [1998], which states that for any prior π, with proba greater than $1-\epsilon$, for any posterior ρ, we have the following bound :

$$\left|\mathbb{E}_{f\sim\rho}R(f) - \mathbb{E}_{f\sim\rho}\widehat{R}(f)\right| \leq \sqrt{\frac{\log(4n\epsilon^{-1}) + \mathcal{K}(\rho,\pi)}{2n-1}}.$$

Then, minimizing the RHS leads to a Gibbs measure ρ by the Kullback duality formula.

Seeger [2008] proposes a simple proof of this inequality. It is based on the convex duality formula (see equation (4.9) in Chapter 4). However, similar duality formula arises for more general divergence such as Bregman divergences defined as :

$$\mathcal{D}_\Phi(\rho,\pi) = \Phi(\rho) - \Phi(\pi) - \langle \rho - \pi, \nabla\Phi(\pi)\rangle,$$

where Φ is a strictly convex and differentiable function and $\langle\cdot,\cdot\rangle$ is the scalar product (such as the negative entropy $\Phi(\mathbf{p}) = \sum_{j=1}^p p_j \log p_j$ for the Kullback-Leibler divergence). Using such a divergence and the general convex duality argument (see Rockafellar [1970]), we can conjecture that for any prior π, with proba greater than $1-\epsilon$, for any posterior ρ :

$$\left|\mathbb{E}_{f\sim\rho}R(f) - \mathbb{E}_{f\sim\rho}\widehat{R}(f)\right| \leq \sqrt{\frac{\Psi(\epsilon^{-1}) + \mathcal{D}_\Phi(\rho,\pi)}{\gamma_n}},$$

where Ψ comes from the convex duality function of $\rho \mapsto \mathcal{D}(\rho,\cdot)$ and γ_n is an increasing sequence with n. This could lead to other randomized sequential procedure such as for instance polynomial weighted averages.

Open problem 16: We can extend the result of Section 4.2 to the standard i.i.d. case by considering a mirror averaging as it is proposed in Section 4.1. We can also consider the high dimensional setting, where $x_t \in \prod_{j=1}^p \mathbb{R}^{m_j}$, with $m_j >> T$. In this case, using a slightly modified prior, we can get a sparsity regret bound where the sparsity is measured thanks to the standard ℓ_0-norm with respect to the coordinate of each cluster's center. This framework could be also investigated in high dimensional sequential clustering. In this case, the number of clusters has to be fixed in the algorithm as a small value (in comparison with T). An open problem is to propose at the same time model selection clustering and high dimensional clustering.

Open problem 17: The algorithm of Section 4.2 could be seen as an alternative of standard kernel-based method in local regression. In Chapter 3, we consider a local M-estimator of the regression function at a fixed point x_0 as :

$$\hat{f}_h(x_0) = \arg\min \frac{1}{n} \sum_{i=1}^{n} Y_i \mathcal{K}_h(X_i - x_0),$$

where $\mathcal{K}_h$ is a kernel function that gives more weights to the points at a neighborhood of x_0. Instead, we can consider a similar local estimator but where localization is performed based on a pre-processing clustering. For this purpose, in the i.i.d. case, we can introduce a mirror averaging estimator based on the sequential procedure of Section 4.2. In this case, localization takes into acccount the structure of the data points. It could be useful in a high dimensional setting where kernel estimators are prohibited.

Open problem 18: The computation of algorithms presented in Chapter 4 remains an hard issue. Recent works has been proposed in the literature of prediction with sparse single index models (see Alquier and Biau [2013]), sparse additive model (see Alquier and Guedj [2013]) or high dimensional linear regression (see Dalalyan and Tsybakov [2012]). In these studies, MCMC methods appears to be efficient to approximate the Gibbs posterior in these sequential algorithms. We are currently implementing these MCMC for sequential clustering but that's another story !

Epilogue

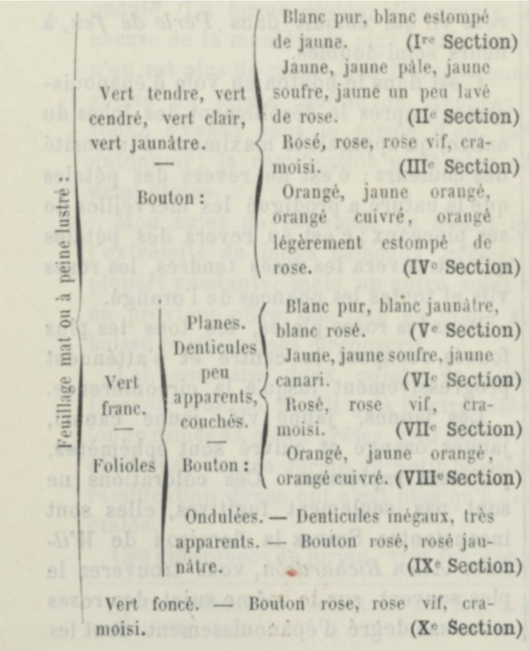
Cet été 2014 fut entièrement dédié à la rédaction de ce mémoire, enfin presque. Comme vous avez pu le remarquer, ce manuscrit se concentre sur le problème de classification (supervisée et non-supervisée). Cette discipline me suit depuis mes premiers pas en thèse, jusqu'aux derniers résultats Pac-Bayésien en prévision de suites individuelles. Cet été 2014 fut donc l'occasion de compiler certains de ces résultats, mais pas seulement. Cela, grâce - ou à cause - de quelques retraités du Val-de-Marne passionnés de rosiers, et plus précisément des rosiers Noisettes.

Nous avons entrepris cet été 2014 un vrai travail de taxinomie, l'essence même de la classification, version botaniste. Une visite de Laurence il y a quelques mois dans mon bureau, munis d'un poster avec 1232 mesures d'inflorescence, rameaux, folioles, épines, pétales, sépales réalisées par les ”Amis de la Roseraie du Val-de-Marne” est à l'origine de ce travail. S'en suit l'encadrement d'un mémoire de M1 sur le sujet, puis ce stage estival avec Charline Bris, embauchée comme ingénieur de recherche sur le projet ”Classification des rosiers noisettes”. Ces quelques lignes synthétisent ce travail de classification.

En 1899, dans le *Journal des roses* d'octobre, Pierre Guillot (1855-1918) écrit :

> Le *Rosa Noisettiana* type est un arbuste dont les rameaux sont nombreux, sarmenteux, buissonnants, terminés par de très grosses ombelles de petites fleurs blanches, roses, pourpres, blanc carné, jaunes, selon les variétés. Parmi les formes qui en sont issues, la plus grande partie est à grandes fleurs, ce qui est le contraire du type ; l'inflorescence n'a plus la même forme et se rapproche beaucoup de celle du *R. Indica.*
>
> A notre avis, cette race *R. Noisettiana* ne devrait comprendre que les variétés du type, comme :
>
> | | |
> |---|---|
> | *Aimée Vibert* | Vibert 1828 |
> | *Bougainville* | P. Cochet 1824 |
> | *Caroline de Marniesse* | Roeser 1848 |
> | *Octavie* | Vibert 1845 |
> | *Ophirie* | Goubault 1841 |
> | *W.-A Richardson* | V^{e} Ducher 1878 |
>
> [...] Quant aux autres, leur regroupement est tout indiqué : on devra les placer avec les *Rosiers thé à rameaux sarmenteux*, de manière à laisser à chaque série le caractère spécial qui lui est propre.

Dans les numéros de septembre, octobre et novembre 1899 du *Journal des roses*, Simon Sirodot[4], après avoir remarqué que les mêmes variétés cultivés en plusieurs exemplaires ont des différences, notamment de forme, proposose une classification des rosiers Noisette en 10 sections.



Feuillage mat ou à peine lustré :

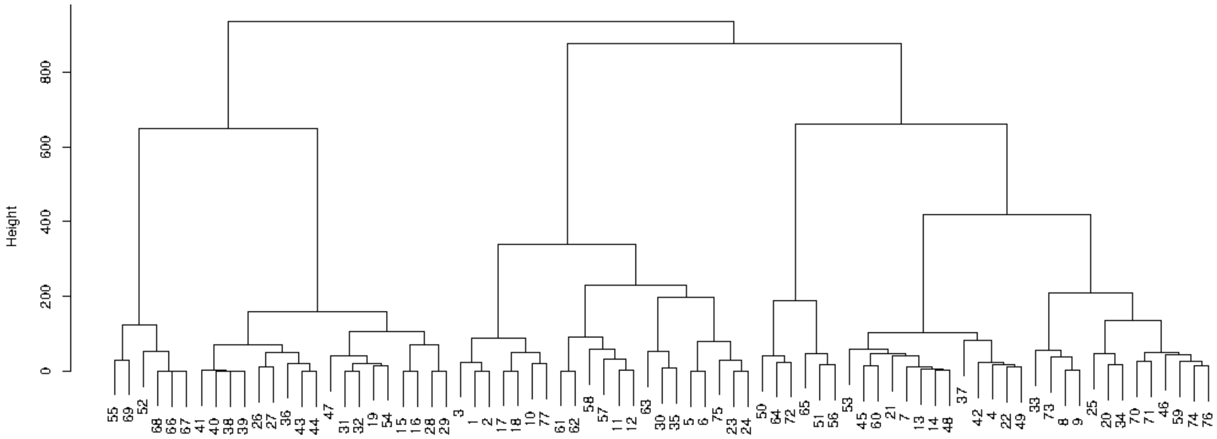
Vert tendre, vert cendré, vert clair, vert jaunâtre. — Bouton :
- Blanc pur, blanc estompé de jaune. (I^re Section)
- Jaune, jaune pâle, jaune soufre, jaune un peu lavé de rose. (II^e Section)
- Rosé, rose, rose vif, cramoisi. (III^e Section)
- Orangé, jaune orangé, orangé cuivré, orangé légèrement estompé de rose. (IV^e Section)

Vert franc. — Folioles
- Planes. Denticules peu apparents, couchés. — Bouton :
  - Blanc pur, blanc jaunâtre, blanc rosé. (V^e Section)
  - Jaune, jaune soufre, jaune canari. (VI^e Section)
  - Rosé, rose vif, cramoisi. (VII^e Section)
  - Orangé, jaune orangé, orangé cuivré. (VIII^e Section)
- Ondulées. — Denticules inégaux, très apparents. — Bouton rosé, rosé jaunâtre. (IX^e Section)

Vert foncé. — Bouton rose, rose vif, cramoisi. (X^e Section)

Figure 1 – Les 10 sections de Sirodot

En s'inspirant de Kuentz-Simonet, Lyser, Candau, Deuffic, Chavent, and Saracco [2012], nous avons effectué plusieurs classifications ascendantes hiérarchiques (CAH) sur notre échantillon de rosiers à partir de variables synthétiquent issues d'une analyse en correspondance multiple. Voici le dendogramme associé à 15 variables synthétiques qu'il reste à interpréter avec "Les Amis de la Roseraie du Val-de-Marne" :



[4] Simon Sirodot (1825-1903) : Membre correspondant de l'Académie des Sciences, section de Botanique et de la Société Nationale d'agriculture, section d'Histoire naturelle.

Bibliographie

[1] Collins A., D. Milbourne, L. Ramsay, R. Meyer, C. Chatot-Balandras, P. Oberhagemann, W. De Jong, C. Gebhardt, E. Bonnel, and R. Waugh. Qtl for field resistance to late blight in potato are strongly correlated with maturity and vigour. *Molecular breeding*, 5 :387–398, 1999.

[2] B. Alamar. *Sport analytics : a guide for coaches, managers, and other decision-makers.* Columbia University Press, 2013.

[3] P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14 :243–280, 2013.

[4] P. Alquier and B. Guedj. Pac-bayesian estimation and prediction in sparse additive models. *Electron. J. Stat.*, 7 : 264–291, 2013.

[5] András Antos, László Györfi, and András György. Individual convergence rates in empirical vector quantizer design. *IEEE Trans. Inf. Theory*, 51(11) :4013–4022, 2005.

[6] Ery Arias-Castro, Joseph Salmon, and Rebecca Willett. Oracle inequalities and minimax rates for nonlocal means and related adaptive kernel-based methods. *SIAM J. Imaging Sci.*, 5(3) :944–992, 2012.

[7] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10 :245–279, 2009.

[8] J. Astola, K. Egiazarian, and V. Katkovnik. Adaptive Window Size Image De-noising Based on Intersection of Confidence Intervals (ICI) Rule. *J. Math. Imaging Vis.*, 16(3) :223–235, 2002. ISSN 0924-9907.

[9] J. Astola, K. Egiazarian, A. Foi, and V. Katkovnik. From Local Kernel to Nonlocal Multiple-Model Image Denoising. *Int. J. Comput. Vision*, 86(1) :1–32, 2010.

[10] J-Y. Audibert. Classification under polynomial entropy and margin assumptions and randomized estimators. Preprint, Laboratoire de Probabilités et Modéles Aléatoires, Univ. Paris VI and VII., 2004.

[11] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37 (4) :1591–1646, 2009.

[12] J-Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35 :608–633, 2007.

[13] P. Auer, N. Cesa-Bianci, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64 :48–75, 2002.

[14] A. Barron. Are bayes rules consistent in information ? In *Open Problems in Communication and Computation*, pages 85–91. Springer New-York, 1987.

[15] A. Barron and G. Leung. Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory*, 52(8) : 3396–3410, 2006.

[16] D. Barthélémy and Y. Caraglio. Plant architecture : a dynamic, multilevel and comprehensive approach to plant form, structure and ontogeny. *Ann. Bot.*, 99 :375–407, 2007.

[17] P.L. Bartlett and S. Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135 (3) :311–334, 2006.

[18] P.L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inf. Theory*, 44(5) :1802–1813, 1998.

[19] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Ann. Statist.*, 33 (4) :1497–1537, 2005.

[20] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics : overview and implementation. *Statistics and Computing*, 22 :455–470, 2012.

[21] G. Biau and A. Fisher. Parameter selection for principal curves. *IEEE Trans. Inf. Theory*, 58, 2012.

[22] G. Biau, L. Devroye, and G. Lugosi. On the performances of clustering in Hilbert spaces. *IEEE Trans. Inf. Theory*, 54 (2), 2008.

[23] P. Bickel and Y. Ritov. Nonparametric estimators which can be "plugged-in". *Ann. Statist.*, 31(4) :1033–1053, 2003. ISSN 0090-5364. doi : 10.1214/aos/1059655904. URL http://dx.doi.org/10.1214/aos/1059655904.

[24] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Ann. Statist.*, 4 :861–894, 2003.

[25] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36 (2) :489–531, 2008.

[26] B.E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152, 1992.

[27] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : a survey of some recent advances. *ESAIM : Probability and Statistics*, 9 :323–375, 2005.

[28] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6) :495–500, 2002.

[29] C. Bouveyron and C. Brunet. Model-based clustering of high-dimensional data : A review. Computational Statistics and Data Analysis, to appear, 2013.

[30] D. Bradley, O. Ratcliffe, C. Vincent, R. Carpenter, and E. Coen. Inflorescence commitment and architecture in arabidopsis. *Science*, 275 :80–83, 1997.

[31] L. Breiman. Statistical modeling : The two cultures. *Statistical Science*, 16 (3) :199–231, 2001.

[32] B.M. Brown. Martingale Central Limit Theorems. *Ann. Math. Statist.*, 42(1) :59–66, 1971.

[33] L. Brown and M. Low. A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.*, 24(6) :2524–2535, 1996.

[34] V.-E. Brunel. Adaptive estimation of convex sets and convex polytopes from noisy data. Preprint, 2013.

[35] S. Bubeck. How the initialization affects the k means. *IEEE Trans. Inf. Theory*, 48 :2789–2793, 2002.

[36] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data.* Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[37] C. Butucea. goodness-of-fit testing and quadratic functionnal estimation from indirect observations. *Ann. Statist.*, 35 :1907–1930, 2007.

[38] P. Calès, J. Boursier, J. Chaigneau, F. Lainé, J. J. Sandrini, S. Michalak, I. Hubert, N. Dib, F. Oberti, S. Bertrais, G. Hunault, C. Cavaro-Ménard, Y. Gallois, Y. Deugnier, and M.C. Rousselet. Diagnosis of different liver fibrosis characteristics by blood tests in non-alcoholic fatty liver disease. *Liver Int.*, 30 (9) :1346–1354, 2010.

[39] P. Calès, J. Boursier, A. Ducancelle, F. Oberti, I. Hubert, G. Hunault, V. Lédinghen, J.P. Zarski, D. Salmon, and F.Lunel. Improved fibrosis staging by elastometry and blood test in chronic hepatitis c. *Liver Int.*, 34 (6) :907–917, 2014.

[40] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9 :717–772, 2009.

[41] O. Catoni. *Statistical Learning Theory and Stochastic Optimization.* Springer, Lecture Notes in Mathematics, 2001.

[42] N. Cesa-Bianchi and G. Lugosi. *Learning, Prediction and Games.* Cambridge University Press, 2006.

[43] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R.E. Schapire, and M. Warmuth. How to use expert advice. *Journal for the Association of Computing Machinery*, 44 (3) :427–485, 1997.

[44] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66 :321–352, 2007.

[45] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik. Vicinal risk minimization. In *Advances in Neural Information Processing Systems*, pages 416–422. MIT Press, 2001.

[46] Y. Cheng and G.M. Church. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2000.

[47] M. Chichignoud. Minimax and minimax adaptive estimation in multiplicative regression : locally Bayesian approach. *Probab. Theory Related Fields*, 153(3-4) :543–586, 2012.

[48] M. Chichignoud and J. Lederer. A Robust, Fully Adaptive M-estimator for Pointwise Estimation in Heteroscedastic Regression. *To appear in Bernoulli, arXiv :1207.4447v3*, 2013.

[49] A. Choromanska and C. Monteleoni. Online clustering with experts. In Journal of Machine Learning Research (JMLR) Workshop and Conference Proceedings, editors, *Proceedings of ICML 2011 Workshop on Online Trading of Exploration and Exploitation 2*, 2012.

[50] G.A. Churchill and R.W. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138 :963–971, 1994.

[51] F. Comte and C. Lacour. Anisotropic adaptive kernel deconvolution. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49 (2) :569–609, 2013.

[52] E. Costes and Y. Guédon. Modelling branching patterns on 1-year- old trunks of six apple cultivars. *Ann. Bot.*, 89 : 513–524, 2002.

[53] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel based learning methods.* Cambridge University Press, 2000.

[54] A. S. Dalalyan and A.B. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *J. Comput. System Sci.*, 78 :1423–1443, 2012.

[55] A.S. Dalalyan and A.B. Tsybakov. Miror averaging with sparsity priors. *Bernoulli*, 18 (3) :914–944, 2012.

[56] I. Dattner, M. Reiß, and M. Trabs. Adaptive quantile estimation in deconvolution with unknown error distribution. *arXiv : 1303.1698*, 2013.

[57] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.

[58] H.W. Engl, M. Hank, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer Academic Publishers Group, Dordrecht, 1996.

[59] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19 :1257–1272, 1991.

[60] J. Fan and Y.K Truong. Nonparametric regression with errors in variables. *Ann. Statist.*, 21 (4) :1900–1925, 1993.

[61] P Ferraro and C Godin. A distance measure between plant architectures. *Ann. For. Sci.*, 57 :445–461, 2000.

[62] P Ferraro, C Godin, and P Prusinkiewicz. Toward a quantification of self-similarity in plants. *Fractals*, 13 :91–109, 2005.

[63] A. Fischer. On the number of groups in clustering. *Statistics and Probability Letters*, 81 :1771–1781, 2011.

[64] R. T. Furbank and M. Tester. Phenomics - technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, 16 :635–644, 2011.

[65] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14 :729–769, 2013.

[66] H. Gibert and P. LeRoy. Comparison of three multitrait methods for qtl detection. *Genet. Sel. Evol.*, 35 :281–304, 2003.

[67] D. Gnatyshak, D. Ignatov, A. Semenov, and J. Poelmans. Analysing online social network data with biclustering and triclustering. In *Proceedings of the Conference on Concept Discovery in Unstructured Data*, pages 30–39. Dmitry I. Ignatov, Sergei O. Kuznetsov, Jonas Poelmans (Eds.), 2012.

[68] C. Godin and Y. Caraglio. A multiscale model of plant topological structures. *J. Theor. Biol.*, 191 :1–46, 1998.

[69] C. Godin, E. Costes, and H. Sinoquet. A method for describing plant architecture which integrates topology and geometry. *Ann. Bot.*, 84 :343–357, 1999.

[70] A. Goldenshluger and O.V. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(3) :1150–1190, 2008.

[71] A. Goldenshluger and O.V. Lepski. Structural adaptation via Lp-norm oracle inequalities. *Probab. Theory and Related Fields*, 143 :41–71, 2009.

[72] A. Goldenshluger and O.V. Lepski. Uniform bounds for norms of sums of independent random functions. *Ann. Probab.*, 39(6) :2318–2384, 2011.

[73] A. Goldenshluger and O.V. Lepski. Bandwidth selection in kernel density estimation : oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3) :1608–1632, 2011.

[74] A. Goldenshluger and A. Nemirovski. On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.*, 6(2) :135–170, 1997.

[75] Larry Goldstein and Karen Messer. Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.*, 20(3) :1306–1328, 1992. ISSN 0090-5364. doi : 10.1214/aos/1176348770. URL `http://dx.doi.org/10.1214/aos/1176348770`.

[76] Y. Guédon, D. Barthélémy, Y. Caraglio, and E. Costes. Pattern analysis in branching and axillary flowering sequences. *J. Theor. Biol.*, 212 :481–520, 2001.

[77] C.A. Hackett, R.C. Meyer, and W.T.B. Thomas. Multi-trait qtl mapping in barley using multivariate regression. *Genet. Res. Camb.*, 77 :95–106, 2001.

[78] C.S. Haley and S.A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69 :315–324, 1992.

[79] P. Hall and S. N. Lahiri. Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.*, 36(5) :2110–2134, 2008.

[80] P.R. Halmos. What does spectral theorem say ? *Amer. Math. Monthly*, 70 :241–247, 1963.

[81] J. Hannan. Approximation to bayes risk in repeated play. In *Contributions to the Theory of Games*, volume III, pages 97–139. Princeton University Press, 1957.

[82] J.A. Hartigan. *Clustering algorithms.* Wiley, 1975.

[83] R. Has'minskii and I. Ibragimov. On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist.*, 18(3) :999–1010, 1990.

[84] R.Z. Has'minskii and I.A. Ibragimov. *Statistical Estimation, Asymptotic Theory.* Springer-Verlag, Applications of Mathematics, 1981.

[85] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, 2002.

[86] D. Haussler, J. Kivinen, and M. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Trans. Inf. Theory*, 44 (5) :1906–1925, 1998.

[87] P.J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35 :73–101, 1964.

[88] H. Iwata, A. Gaston, A. Remay, T. Thouroude, J. Jeauffre, K. Kawamura, L. Hibrand-Saint Oyant, T. Araki, B. Denoyes, and F. Foucher. The tfl1 homologue ksn is a regulator of continuous flowering in rose and strawberry. *Plant J.*, 69 :116–125, 2012.

[89] R.C. Jansen and P Stam. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 136 :1447–1455, 1994.

[90] C Jiang and Z.B Zeng. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140 :1111–1127, 1995.

[91] I.M. Johnstone and B.W. Silverman. Empirical bayes selection of wavelet thresholds. *Ann. Statist.*, 33 :1700–1752, 2005.

[92] A. Juditsky, P. Rigollet, and A.B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36 (5) :2183–2206, 2008.

[93] V. Katkovnik. A new method for varying adaptive bandwidth selection. *IEEE Trans. Image Process.*, 47(9) :2567–2571, 1999.

[94] V. Katkovnik and V. Spokoiny. Spatially adaptive estimation via fitted local likelihood techniques. *IEEE Trans. Signal Process.*, 56(3) :873–886, 2008.

[95] K. Kawamura, L. Hibrand-Saint Oyant, L. Crespel, T. Thouroude, D. Lalanne, and F. Foucher. Quantitative trait loci for flowering time and inflorescence architecture in rose. *Theor Appl Genet*, 122 (4) :661–675, 2011.

[96] G. Kerkyacharian, O.V. Lepski, and D. Picard. Non linear estimation in anisotropic multi-index denoising. *Probab. Theory and Related Fields*, 121 :137–170, 2001.

[97] Ch. Kervrann and J. Boulanger. Optimal Spatial Adaptation for Patch-Based Image Denoising. *IEEE Trans. Image Process.*, 15(10) :2866–2878, 2006.

[98] N. Klutchnikoff. *On the adaptive estimation of anisotropic functions.* Ph.D. thesis, Aix-Masrseille 1, 2005.

[99] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6) : 2593–2656, 2006.

[100] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008.

[101] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*, pages 443–459, 2000.

[102] V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39 (5) :2302–2329, 2011.

[103] A. P. Korostelëv and A. B. Tsybakov. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993. ISBN 0-387-94028-6.

[104] A.P. Korostelev and A.B. Tsybakov. *Minimax theory of Image Reconstruction. Lecture Notes in Statistics.* Springer Verlag, 1993.

[105] S. Kotz and S. Nadarajah. *Multivariate t distribution and their applications.* Cambridge University Press, 2004.

[106] V. Kuentz-Simonet, S. Lyser, J. Candau, P. Deuffic, M. Chavent, and J. Saracco. Classification de variables qualitatives pour la compréhension de la prise en compte de l'environnement par les agriculteurs. XIèmes Journées de Méthodologie Statistique de l'Insee, 2012.

[107] E.S. Lander and D Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121 :185–199, 1989.

[108] N. B. Langlade, X. Feng, T. Dransfield, L. Copsey, A. I. Hanna, C. Thébaud, A. Bangham, A. Hudson, and E. Coen. Evolution through genetically controlled allometry space. *PNAS*, 102 :10221–10226, 2005.

[109] G. Lecué and S. Mendelson. General non-exact oracle inequalities for classes with a subexponential envelope. *Ann. Statist.*, 40 (2) :832–860, 2012.

[110] Guillaume Lecué. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, 35(4) : 1698–1721, 2007.

[111] Y. Lederer and S. van de Geer. New concentration inequalities for suprema of empirical processes. Submitted, 2012.

[112] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness : an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3) :929–947, 1997.

[113] O.V. Lepski. On a Problem of Adaptive Estimation in Gaussian White Noise. *Theory Prob. App.*, 35(3) :454–466, 1990.

[114] O.V. Lepski. Asymptotically minimax adaptive estimation I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.*, 36 :682–697, 1991.

[115] O.V. Lepski. Asymptotically minimax adaptive estimation II. Statistical models without optimal adaptation. Adaptive estimators. *Theory Prob. App.*, 37 :433–468, 1992a.

[116] O.V. Lepski. On problems of adaptive estimation in white Gaussian noise. *In Topic in Nonparametric Estimation*, 12 :87–106, 1992b.

[117] C. Levrard. Fast rates for empirical vector quantization. *Electron. J. Stat.*, 7 :1716–1746, 2013.

[118] N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108 :212–261, 1994.

[119] S.P. Lloyd. Least square quantization in pcm. *IEEE Trans. Inf. Theory*, 28 (2) :129–136, 1982.

[120] C. X. Ma, G. Casella, and R. L. Wu. Functional mapping of quantitative trait loci underlying the character process : a theoretical framework. *Genetics*, 161 :1751–1762, 2002.

[121] D.A. Mac Allester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

[122] S. Mallat. *A wavelet tour of signal processing.* Elsevier/Academic Press, Amsterdam, third edition, 2009. The sparse way, With contributions from Gabriel Peyré.

[123] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6) :1808–1829, 1999.

[124] K.V. Mardia, J.T. Kent, and J.M. Bibby. Discriminant analysis. *Multivariate Analysis, Academic Press, London*, pages 300–332, 1979.

[125] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9 (2) : 245–303, 2000.

[126] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour.

[127] P. Massart and E. Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34 :2326–2366, 2006.

[128] P. Mathé. The Lepskiĭprinciple revisited. *Inverse Problems*, 22(3) :L11–L15, 2006.

[129] R. C. McGarry and B. G. Ayre. Manipulating plant architecture with members of the cets gene family. *Plant Sci.*, 188-189 :71–81, 2012.

[130] A. Meister. *Deconvolution problems in nonparametric statistics*, volume 193 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 2009.

[131] S. Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4 :759–771, 2003.

[132] Karl J. Niklas. *Plant allometry : The scaling of form and process.* University of Chicago Press, 1994.

[133] S. M. Nikol'skii. *Approximation of functions of several variables and imbedding theorems.* Springer-Verlag, New York, 1975.

[134] K. Onishi, Y. Horiuchi, N. Ishigoh-Oka, K. Takagi, N. Ichikawa, M. Maruoka, and Y. Sano. A qtl cluster for plant architecture and its ecological significance in asian wild rice. *Breed. Sci.*, 57 :7–16, 2007.

[135] J. W. Van Ooijen. *MAPQTL® 5.0 software for the mapping of quantitative trait loci in experimental populations.* Plant Research International, Wageningen, 2004.

[136] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data : A review. *SIGKDD Explor. Newsl.*, 6(1) :90–105, June 2004. ISSN 1931-0145.

[137] E. Parzen. On the estimation of a probability density function and mode. *Ann. Math. Statist.*, 33 :1065–1076, 1962.

[138] R. W. Pearcy, H. Muraoka, and F. Valladares. Crown architecture in sun and shade environments : assessing function and trade-offs with a three-dimensional simulation model. *New Phytol.*, 166 :791–800, 2005.

[139] D. Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9(1) :135–140, 1981.

[140] D. Pollard. A central limit theorem for k-means clustering. *Ann. Probab.*, 10(4) :919–926, 1982.

[141] W. Polonik. Measuring mass concentrations and estimating density contour clusters - an excess mass approach. *Ann. Statist.*, 23 (3) :855–881, 1995.

[142] J. Polzehl and V. Spokoiny. Propagation-separation approach for local likelihood estimation. *Probab. Theory Related Fields*, 135(3) :335–362, 2006.

[143] P. Prusinkiewicz, Y. Erasmus, B. Lane, L. D. Harder, and E. Coen. Evolution and development of inflorescence architectures. *Science*, 316 :1452–1456, 2007.

[144] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning : Random averages, combinatorial parameters, and learnability. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1984–1992. Curran Associates, Inc., 2010.

[145] M. Renton, Y. Guédon, C. Godin, and E. Costes. Similarities and gradients in growth unit branching patterns during ontogeny in fuji apple trees : a stochastic approach. *J. Exp. Bot.*, 57 :3131–3143, 2006.

[146] J. Resig. *Pro Javascript Techniques.* Apress, 2006.

[147] V. Rivoirard. Nonlinear estimation over weak besov spaces and minimax bayes method. *Bernoulli*, 12 :609–632, 2006.

[148] R.T. Rockafellar. *Convex analysis.* Princeton Landmarks in Mathematics, 1970.

[149] M. Rosenblatt. Remarks on som nonparametric estimates of a density function. *Ann. Math. Statist.*, 23 :832–837, 1956.

[150] T. Sakamoto and M. Matsuoka. Generating high-yielding varieties by genetic manipulation of plant architecture. *Current Opinion in Biothechnology.*, 15 :144–147, 2004.

[151] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, 2002.

[152] B. Scholkopf, K. Tsuda, and J.-P. Vert. *Kernel methods for computational biology.* MIT, 2007.

[153] M.W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9 :759–813, 2008.

[154] V. Segura, C. Cilas, F. Laurens, and E. Costes. Phenotyping progenies for complex architectural traits : a strategy for 1-year-old apple trees (malus x domestica borkh.). *Tree Genetics and Genomes*, 2 :140–151, 2006.

[155] V. Segura, A. Ouangraoua, P. Ferraro, and E. Costes. Comparison of tree architecture using tree edit distances : application to 2-year-old apple hybrids. *Euphytica*, 161 :155–164, 2008.

[156] Y. Seldin. A pac-bayesian approach to structure learning. Phd Thesis, The Hebrew University of Jerusalem, 2009.

[157] Y. Seldin and N. Tishby. Pac-bayesian analysis of co-clustering and beyong. *Journal of Machine Learning Research*, 11 :3595–3646, 2010.

[158] N. Silver. *Baseball Between the Numbers : Why Everything You Know About the Game Is Wrong.* Jonah Keri Editions, 2007.

[159] N. Slonim and N. Tishby. Document clustering using word clusters via the information botleneck method. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrievial*, 2000.

[160] X. Song and T. Zhang. Quantitative trait loci controlling plant architectural traits in cotton. *Plant Sci.*, 177 :317–323, 2009.

[161] V. Spokoiny and C. Vial. Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, 37 (5B) :2783–2807, 2009.

[162] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S*, 81 :73–205, 1995.

[163] F Tian, P.J. Bradbury, P.J. Brown, H Hung, Qi Sun, S Flint-Garcia, T. R. Rocheford, M. D. McMullen, J. B. Holland, and E. S. Buckler. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics*, 43 :159–162, 2011.

[164] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society*, 63 :411–423, 2001.

[165] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1) :135–166, 2004.

[166] A.B. Tsybakov. *Introduction à l'estimation non-paramétrique.* Springer-Verlag, 2004.

[167] A.B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009.

[168] A.B. Tsybakov and S.A. van de Geer. Square root penalty : adaptation to the margin in classification and in edge estimation. *Ann. Statist.*, 33 (3) :1203–1224, 2005.

[169] N. Upadyayula, J. Wassom, M.O. Bohn, and T.R. Rocheford. Quantitative trait loci analysis of phenotypic traits and principal components of maize tassel inflorescence architecture. *Theor Appl Genet*, 113 :1395–1407, 2006.

[170] S. van de Geer. *Empirical Processes in M-estimation.* Cambridge University Press, 2000.

[171] A. W. van der Vaart and J. A. Wellner. *Weak convergence and Empirical Processes. With Applications to Statistics.* Springer Verlag, 1996.

[172] V. Vapnik. *The Nature of Statistical Learning Theory.* Statistics for Engineering and Information Science, Springer, 2000.

[173] V. N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[174] V. N. Vapnik and A. Y. Chervonenkis. Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data. *Avtomat. i Telemeh.*, (2) :42–53, 1971.

[175] U. von Luxburg, R. Williamson, and I. Guyon. Clustering : Science or art ? Opinion paper for the NIPS workshop Clustering : Science or Art, 2009.

[176] V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69 (2) :213–248, 2001.

[177] J.I. Weller, G.R. Wiggans, P.M Vanraden, and M. Ron. Application of canonical a transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. *Theor Appl Genet*, 92 :998–1002, 1996.

[178] R. Wu, J. Cao, Z. Huang, Z. Wang, J. Gai, and E. Vallejos. Systems mapping : how to improve the genetic mapping of complex traits through design principles of biological systems. *BMC Systems Biology*, 5 :84, 2011.

[179] R. L. Wu. Genetic mapping of qtls affecting tree growth and architecture in populus : implication for ideotype breeding. *Theoretical and Applied Genetics*, 96 :447–457, 1998.

[180] Y. Yang. Minimax nonparametric classification—part i : Rates of convergence. *IEEE Trans. Inf. Theory*, 45 : 2271–2284, 1999.

[181] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68 (1) :49–67, 2007.

[182] F. Zhang, J. Jiang, S. Chen, F. Chen, and W. Fang. Mapping single-locus and epistatic quantitative trait loci for plant architectural traits in chrysanthemum. *Mol. Breed.*, 30 :1027–1036, 2012.

[183] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D.N. Metaxas. Automatic image annotation using group sparsity. In *In CVPR*, 2010.

[184] X.B. Zhou, C. Chen, Z.C. Li, and X.Y Zhou. Using chou's amphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J Theor Biol*, 248 :546–551, 2007.

[185] S. Zong. Efficient online spherical k-means clustering. In IEEE, editor, *Proceedings of International Joint Conference on Neural Networks IJCNN'05*, pages 3180–3185, 2005.

Index